FIRST PRINCIPLES STUDY OF INTEIN REACTION MECHANISMS

By

Philip Shemella

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: PHYSICS

Approved by the
Examining Committee:

_____
Saroj K. Nayak, Thesis Adviser

_____
Georges Belfort, Member

_____
Marlene Belfort, Member

_____
Angel E. García, Member

_____
Shekhar Garde, Member

_____
György Korniss, Member

Rensselaer Polytechnic Institute
Troy, New York

May 2008
(For Graduation May 2008)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I know but one freedom and that is the freedom of the mind.
– Antoine de Saint Exupéry

ABSTRACT

Protein splicing is an autocatalytic reaction where two flanking sequences (exteins) are excised and ligated. The enzymatic protein sequence that lies between the exteins, known as the intein, is extremely efficient at protein splicing and has been utilized for biotechnological applications. The characterization of intein reaction mechanisms is important for understanding how and why certain mutations may be used to control the splicing and cleavage reactions, as well as tuning the reaction rate without affecting the mechanism. In combination with crystallographic structures as well as both site-directed and random mutagenesis, we have studied the reaction mechanisms for intein splicing as well as for cleavage at the C-terminus using first principle quantum mechanical simulations. Previous experimental studies have shown that mutation at a critical N-terminal residue of the intein resulted in splicing inhibition. Despite this inhibition of the overall splicing reaction, peptide bond cleavage isolated at the C-terminal may still occur independently. With an aspartate to glycine mutation, the "cleavage mutant" was found to react more rapidly in a low pH environment. We have characterized the pH dependent C-terminal cleavage reaction and studied the effect of mutation on the energy barrier, and have provided for the first time an atomic level understanding of this important process. Next, we have extended our computational study to address the overall intein splicing mechanism. The splicing reaction is a highly synchronized chemical process where the effect of mutation can accelerate, decelerate, or partially or completely inhibit steps along the reaction. We have focused our study on the energetic effect of mutation on the reaction profile and corresponding protein structure. From this, we have made a prediction for the splicing mechanism that utilizes the highly conserved amino acids and explicitly describes the behavior of protons. An explanation for the experimental inhibition of splicing with the aspartate to glycine mutation is also presented. In summary, with a series of quantum mechanical calculations ranging from gas phase, to an implicit solvent scheme, to combined quantum/classical simulations, we have provided insight into some of the key steps of intein reactions. These studies may be exploited for many applications involving inteins including molecular switches and sensors as well as controlled drug delivery.

CHAPTER 1
INTRODUCTION

1.1 Computer simulations in chemistry

Nestled between experiment and pure theory, computational chemistry has become an integral tool for researchers working in physics, chemistry, and biology, as well as nanotechnology and biotechnology. The role of simulation is to verify and confirm or to predict and suggest experimental studies. Computer simulations allow the researcher to access states both visible and invisible to experiment, and make predictions based on this knowledge. A chemical reaction may be quantified by the amount of reactants, the amount of products, and the time elapsed. To explain a mechanism and molecular structure and energies on the atomic level, computational methods are important.

The field of computational chemistry spans length and time scales, and the desired level of accuracy is important to know prior to running simulations. To simulate protein folding, which requires an extremely long simulation trajectory, amino acids may be "coarse-grained," where the atomic description of each side chain is aggregated into a composite value. To achieve long trajectories this approximation as well as others are essential. However, to calculate the pK_a of a side chain or the chemical shifts *via* nuclear magnetic resonance (NMR), not only will an atomic level description be necessary, but also a method that can calculate observable properties from first principles is required.

1.2 Computational quantum mechanics

The energies associated with bond breakage and formation are an essential property for an enzymatic processes. For example, a change in energy barrier of ~1.4 kcal/mol corresponds to an order of magnitude change in the reaction rate. States observed at equilibrium may be predicted based on relative energies between structures. To computationally access the energy of the system, and to do so not only for equilibrium structures but also for transition states, first principles electronic structure calculations are required. Using an all-electron method, the electron orbitals are considered variable and flexible, and they depend on neighboring atoms and environment. This is important because the chemistry at transition states may vary greatly from equilibrium structures: instead of four

bonds, carbon atoms may have three or five bonds during a chemical reaction. Transition states are where quantum mechanical principles dominate. By solving the Schrödinger equation for all electrons, and relaxing their orbital positions and therefore allowing the electron density to vary, an accurate description of the system can be obtained that is useful for understanding fundamental chemistry both near and far from equilibrium.

1.3 Inteins

1.3.1 Overview

Protein splicing involves the autocatalytic release of a peptide segment, termed an intein, with the joining of two flanking protein sequences (exteins) [1, 2]. Inteins are autocatalytic proteins that exist in all three domains of life. Experiments have identified key reaction steps in protein splicing whereas sequence comparisons have revealed the conserved amino acids required for this reaction. Figure 1.1 shows a schematic for conserved intein residues and their corresponding block (C or N) designation. Experimental mutational studies have been carried out to further control the protein splicing reaction [3, 4]. For example, by mutating the first residue at the N-terminus (N1 block) of the intein from Cys to Ala (N1-Cys1Ala), the first step of the splicing reaction, namely the N-terminal N–S shift[1], is inhibited, thus isolating the C-terminal cleavage reaction [5]. Mutation schemes that control the reaction rate and/or the specific products could be exploited in many biotechnological applications such as bioseparations [6, 7], drug development [8], and molecular sensors [9, 10].

1.3.2 Computational methodology

First principles density functional theory was used to study the electronic structure of protein systems. Various methods were utilized for accurate description of the intein model system: the electrostatic environment was either neglected (gas phase), approximated (implicit solvent), or treated explicitly (QM/MM). Various levels of theory were used: classical molecular mechanics, semi-empirical quantum mechanics, and first principles quantum mechanics.

---

[1]Atoms are annotated with one letter, *i.e.* H = hydrogen. Amino acids are annotated with three letters, *i.e.* His = histidine.
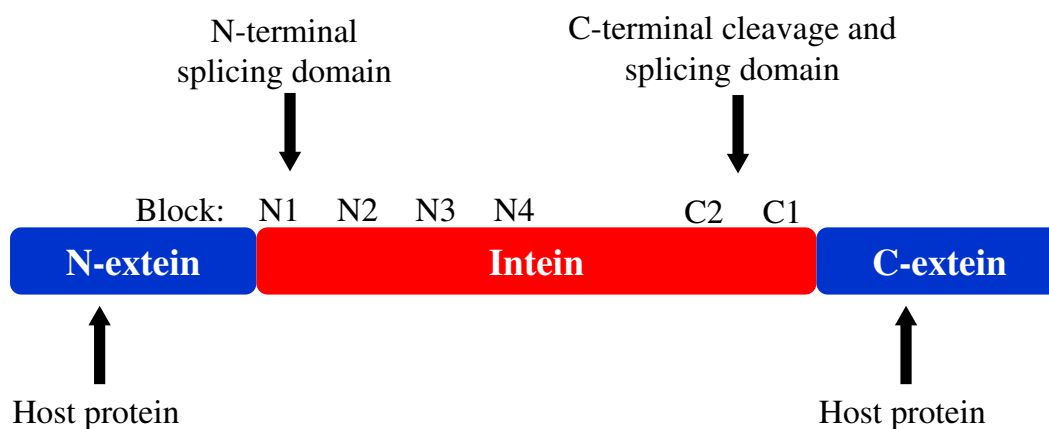
Figure 1.1: Schematic intein and N- and C-exteins. Splicing motifs contain highly conserved amino acids, such as N1-Cys1, N3-His10, C2-Asp5, and C1-His7, C1-Asn8, C1-Cys+1
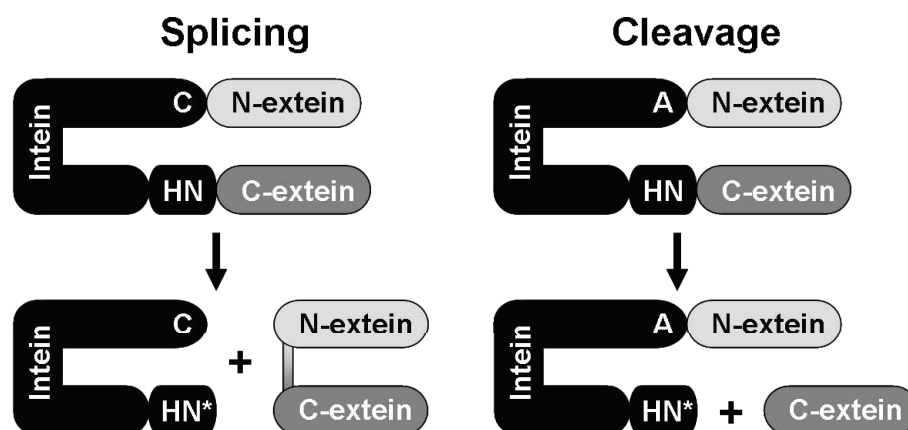
Figure 1.2: Intein reactions: splicing and cleavage (after N1-Cys1Ala mutation). C, A, H, N, and N* represent Cys, Ala, His, Asn, and succinimide, respectively

1.3.3   Results: Splicing

We have studied the protein splicing mechanism for inteins. The role of mutations for the *Mtu* recA intein is considered, especially the C2-Asp5Gly mutation that inhibits splicing and creates the C-terminal cleavage mutant (CM). A mechanism is proposed that is consistent with crystal [11, 12, 13, 14, 15, 16, 17] or NMR structures [18] and mutagenesis results [19, 20, 21, 22, 23], and includes an atomic-level description of the steps of intein splicing. From this mechanism we can gain a detailed understanding of a reaction that is useful for biotechnology applications.

Experiments have identified key reaction steps in protein splicing as well as revealed the conserved amino acids required for this reaction. The intein splicing reaction in terms of conserved amino acids is shown in Figure 1.2. Intein splicing starts with an N-S shift and thioester formation at the N-terminal N1-Cys1 (see Appendix A for all natural amino acid structures). Next is transesterification, where the C-terminal Cys+1 attacks the carbon atom of the N-terminal thioester. Following that, C1-Asn8 cyclization occurs and finally the S-N shift leads to separate products of fused N- and C-exteins and the released intein sequence.

1.3.4   Results: C-terminal cleavage

The cleavage of the peptide bond between the intein and the C-extein was investigated. This is one step that was isolated by inhibiting other steps in the overall splicing reaction, and was found to occur more rapidly in a low pH environment. Our first principles calculations provide an atomic-level description of C-terminal cleavage that uses a hydronium ion to include a pH–induced reaction. Various model systems are used and results from these calculations are compared with experiment with good agreement.

Experimental mutagenesis studies have been carried out to further control the protein splicing reaction (see Figure 1.1 and Appendix B for intein sequences); for example, by mutating the first residue at the N-terminus of the intein from N1-Cys1Ala, the first step of the splicing reaction, namely the N-terminal N-S shift, is inhibited, thus isolating the C-terminal cleavage reaction (see Figure 1.3) [5].

1.3.4.1   Environmental effects of single amino acid mutation

In a related context, experimental results have shown for the *Mycobacterium tuberculosis* (*Mtu*) recA mini-intein that the C2-Asp5Gly mutation[2] creates the highly active C-terminal cleavage mutant (CM) that was experimentally reported to be more active in a low pH environment [19, 24, 25, 26].

Interestingly, once splicing is inhibited, the C-terminal Cys+1 residue (which is the first amino acid of the C-terminal extein or C-extein) is found to be functionally unnecessary. Wood *et al.* have found that this amino acid regulates the reaction rate but does not alter the mechanism [24]. Furthermore, since the CM is found to be exceedingly reactive in a low pH environment, they have utilized Met, which is the native N-terminus

---

[2]C2-Asp5 indicates aspartate at the fifth position of the C2-block. C2-Asp5Gly indicates that aspartate is mutated to glycine. This mutation was previously referred to as D422G.
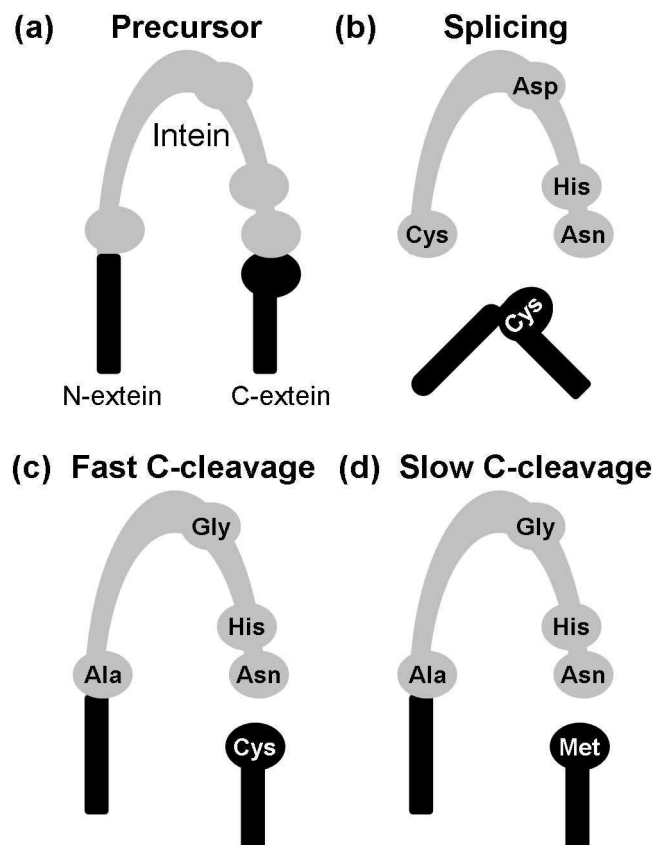
Figure 1.3: Different reaction rates and reaction products are observed for various mutations to the intein or extein sequence. Intein and extein precursor (top left) and two possible products: splicing product (top right), and fast (bottom left) and slow (bottom right) C-terminal cleavage.

of the C-extein sequence, to slow down the reaction by an order of magnitude. In this experiment, three proteins of various sizes were analyzed with the Cys(+1)Met C-extein mutation: Thymidylate Synthase (31.5 kDa), Hfq Protein (18 kDa), and rh aFGF (14 kDa). For these proteins the ratios of reaction rates between the Cys(+1)Met mutants are found to be 12.0, 5.0, and 7.86, respectively [24]. Figure 1.3 shows a schematic of the intein precursor and products based on these results [10, 27], although the exact mechanisms that govern the splicing and cleavage reactions are not understood at the atomic level.

In order to obtain an atomic-level understanding on the reaction mechanisms as well as on the effect of mutation on the reaction barrier, we have carried out detailed quantum mechanical simulations on intein C-terminal cleavage reactions. We describe

pH dependent C-terminal cleavage calculations for the *Mtu* recA intein; performed with semi-empirical, QM gas phase, QM implicit solvent, and combined QM/MM calculations [28, 29]. Harnessing the C-terminal cleavage reaction may allow for an intein-based delivery device, where the reaction is triggered by a certain stimulus.

1.4 Research outline

The goal of our research is to characterize intein splicing and cleavage reaction mechanisms for use in biotechnology applications, and to compliment the experimental efforts of collaborators at RPI and the Wadsworth Center. Understanding the splicing and cleavage mechanisms on the atomic level will provide input for the engineering of smaller and faster intein reactions, as well as a controllable reaction. In addition, the possibility of a synthetic intein is appealing for biotechnological application due to the ability to control parameters such as size, reaction, and function [30].

Our computational results indicate that certain mutations either inhibit or enhance specific reaction steps of the overall splicing reaction, a conclusion that is consistent with experiment. With quantum mechanical simulations, intermediate states may be isolated and studied in the context of altering the molecular triggers and inhibitors that impact protein splicing with inteins. The ability to study precursor, intermediate, and post-reaction product states is extremely useful and carried out with first principles methods.

CHAPTER 2
COMPUTATIONAL METHODS

2.1   Introduction

The field of computational chemistry is extremely broad and includes many methods that encompass a wide range of length and time scales. For example, to simulate large-scale protein structural rearrangement, each atom can be quantized as a partial point charge centered on the atom. By calculating forces from the potential energy of the 'sea' of point charges, and then by integrating Newton's equations of motion, atomic trajectories over long time scales are achievable. One issue with parameterization of the nucleus and the electrons is that once the parameters are set, typical calculations do not include the possibility of "on the fly" re-parameterization. To polarize the partial charge or to break or form new covalent bonds adds an additional level of complexity. In enzyme catalysis, the breaking and forming of bonds is critical, and during these reactions the atoms pass from energetically stable states through the transition states; entering a new stable state. The chemistry involved in the intermediate states is atypical (*i.e.* C atoms may have more or less than four bonds), hence the elevated energy of the system. Figure 2.1 shows a sample reaction energy profile for enzyme catalysis.

2.2   Many-body quantum mechanics

In order to understand both the geometric and electronic structure at the intermediate states, which is essential for accurate prediction of energy barriers and reaction rates, we have used computational quantum mechanics. With these first principles calculations, the all-electron wave function is optimized in a self-consistent manner in order to predict chemical properties and structures. Approximate solutions to the many-electron system will be discussed first by starting with the single electron Hamiltonian.

2.2.1   One-electron Hamiltonian

For non-interacting electrons, we can write the total system Hamiltonian as the sum of one-electron Hamiltonians:

$$H = \sum_{i=1}^{N} h_i \tag{2.1}$$
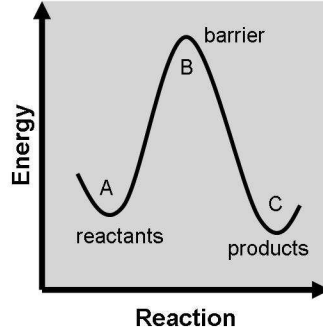
Figure 2.1: Illustrative energy barrier for an enzymatic process (one dimensional reaction coordinate). The difference in energy between the reactants (E_A) and the barrier (E_B) is known as the energy of activation, ΔE^a, and governs the rate of reaction. The difference in energy between the reactants and the products (E_C) may be positive (endothermic) or negative (exothermic), which is useful for predicting which state (reactants or products) is more likely at equilibrium.

where N is the total number of electrons. The classical one-electron Hamiltonian is given by

$$h_i = -\nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} \tag{2.2}$$

which is composed of the electron kinetic energy and the nuclear-electron attraction (M is the total number of nuclei). Eigenfunctions ψ_i must satisfy

$$h_i \psi_i = \epsilon_i \psi_i \tag{2.3}$$

which is the one-electron Schrödinger equation. The many-electron eigenfunction is then the product of the one-electron eigenfunctions

$$\Psi = \prod_{i=1}^{N} \psi_i \tag{2.4}$$

Because each one-electron Hamiltonian (h_i) acts only on its corresponding eigenfunction (ψ_i), the overall system Hamiltonian may be written as

$$H\Psi = \sum_{i=1}^{N} h_i \prod_{i=1}^{N} \psi_i = (\sum_{i=1}^{N} \epsilon_i)\Psi \tag{2.5}$$

2.2.2 The Born-Oppenheimer approximation

Protons and electrons experience an electrostatic force on the same order of magnitude due to an equal magnitude of charge. Because of this, the change in their momenta must also be the same. The mass of a nucleon is 1836 times that of an electron, and from this, the nuclei (which consists of many nucleons) will have a much smaller velocity compared an the electron. On the time scale of nuclear motion, we can consider that the electrons will relax to the ground state electronic structure. The total wave function for nuclei and electrons, $\tilde{\Psi}$ can be separated into the electronic wave function, Ψ, and the nuclear wave function Φ:

$$\tilde{\Psi}(\{\mathbf{r}_i\}, \{\mathbf{r}_\alpha\}) = \Psi(\{\mathbf{r}_i\}, \{\mathbf{r}_\alpha\})\Phi(\{\mathbf{r}_\alpha\}) \tag{2.6}$$

The Born-Oppenheimer approximation treats the nuclei as classical and stationary particles, and the electrons are said to on the Born-Oppenheimer surface. Nuclear kinetic energy is neglected when determining the electronic structure.

This approximation does not work for light nuclei at low temperatures, where a quantum mechanical description of the nuclei is required and usually considered by path integral Monte Carlo (PIMC) or molecular dynamics (PIMD). For the purposes of our biological systems at standard pressure and biological temperatures, the Born-Oppenheimer approximation is sufficient for accurate electronic structure calculations.

2.3 Density functional theory

First principles density functional theory (DFT) [31, 32] has been used in our study of intein splicing and cleavage mechanisms. In this all-electron approximation to the Schrödinger equation, the electron density can be expressed in terms of the wave function:

$$\rho(\mathbf{r}) = |\psi|^2, \tag{2.7}$$

and from the density the number of electrons can be determined:

$$N = \int \rho(\mathbf{r})d\mathbf{r}. \tag{2.8}$$

DFT allows for the ground state properties of a chemical system to be determined without considering the many-electron wave function. Instead, N electrons are grouped into

one composite density, which is determined in a static potential of nucleus-electron and electron-electron interactions.

2.3.1 Classical beginnings

So far, we have written the one-electron Hamiltonian for non-interacting electrons in a lattice of nuclei (Equation 2.2). To add explicit electron-electron interactions, the energy is separated into potential and kinetic terms; starting from classical energy equations and eventually adding quantum terms. The classical potential energy between nuclei and electrons (V_{ne}) and between electrons and electrons (V_{ee}) are given by the functionals of the density $\rho(\mathbf{r})$:

$$V_{ne}[\rho(\mathbf{r})] = \sum_k^{nuclei} \int \frac{Z}{|\mathbf{r} - \mathbf{r}_k|} \rho(\mathbf{r}) d\mathbf{r} \tag{2.9}$$

$$V_{ee}[\rho(\mathbf{r})] = \frac{1}{2} \int \int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2. \tag{2.10}$$

The classical kinetic energy of non-interacting electrons is used:

$$T_{ni} = -\frac{1}{2} \nabla_i^2. \tag{2.11}$$

2.3.2 Quantum corrections to classical terms

The actual kinetic energy is difficult to quantify due to electron correlation effects: specifically, we have not explicitly calculated the correlation interaction due to wave functions for excited electronic states. But, there does exist some electron correction to this classical kinetic energy which we call the correlation energy, E_c. Also, there are the exchange forces due to spin, which have not been included. These interactions are called the exchange energy E_x and can be written exactly as

$$E_x = \int d\mathbf{r}_1 d\mathbf{r}_2 \psi_a^\star(\mathbf{r}_1) \psi_b(\mathbf{r}_1) r_{12}^{-1} \psi_b^\star(\mathbf{r}_2) \psi_a(\mathbf{r}_2). \tag{2.12}$$

The electron exchange arises from the Pauli exclusion principle for fermions. The total wave function for a fermionic system, which is the tensor product of the spatial wave function and the spin wave function, must be anti-symmetric. If two electrons are in the same spatial state (*i.e.* 1s orbital), then their spin wave function must be antiparallel. From this, there is a force, the exchange force, that keeps electrons with opposite spin

in the same spatial orbital. Likewise, electrons with parallel spin experience a force that keeps them from entering the same spatial orbital.

The exact exchange shown in Equation 2.12 is rarely calculated explicitly with current DFT methods due to computational expense of the integration, so the these two quantum mechanical corrections are combined to be the exchange-correlation energy E_{xc}, which is cleverly defined as the difference between the actual total energy and the classical total energy. Feynman called the exchange-correlation energy the "stupidity energy" due to our inability to solve it exactly for non-arbitrary systems [33].

$$E_{xc}[\rho(\mathbf{r})] = E_{total}[\rho(\mathbf{r})] - T_{ni}[\rho(\mathbf{r})] - V_{ne}[\rho(\mathbf{r})] - V_{ee}[\rho(\mathbf{r})] \tag{2.13}$$

A great deal of theoretical effort has been and continues to be spent on creating exchange-correlation functionals that are useful for a broad range of chemical systems.

2.3.3 Hohenberg-Kohn theorems

As a result of the Born-Oppenheimer approximation, the Coulomb potential of the nuclei can be considered a static external potential:

$$V_{ext}(\mathbf{r}) = - \sum_k^{nuclei} \int \frac{Z}{\mid \mathbf{r} - \mathbf{r}_k \mid} d\mathbf{r}. \tag{2.14}$$

The Hamiltonian for electrons can then be written as

$$\hat{F} = -\frac{1}{2}\sum_i \nabla_i^2 + \frac{1}{2}\sum_i \sum_{j \neq i} \frac{1}{\mid \mathbf{r}_i - \mathbf{r}_j \mid} \tag{2.15}$$

so that tht total Hamiltonian is

$$\hat{H} = \hat{F} + \hat{V}_{ext}. \tag{2.16}$$

2.3.3.1 Density determines unique potential

Returning to the external potential, it can be shown that it is uniquely determined by the ground state electronic density. This proof, which is by *reductio ad absurdum*, was first shown by Hohenberg and Kohn [31]. The external potential $V_{ext}(\mathbf{r})$ is related to a ground state wave function $|\Psi_0\rangle$ and density $\rho(\mathbf{r})$. Now consider a second external potential $V'_{ext}(\mathbf{r})$ which corresponds to a different ground state wave function $|\Psi'_0\rangle$ but the

same density $\rho(\mathbf{r})$. The ground state energies of the two systems are

$$E_0 = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle \tag{2.17}$$

$$E_0' = \langle \Psi_0' | \hat{H}' | \Psi_0' \rangle. \tag{2.18}$$

We then use $|\Psi_0'\rangle$ as a trial wave function for the Hamiltonian $\hat{H}$. Because $|\Psi_0'\rangle$ is not an eigenfunction for $\hat{H}$, we know that

$$E_0 < \langle \Psi_0' | \hat{H} | \Psi_0' \rangle \tag{2.19}$$

$$E_0' < \langle \Psi_0 | \hat{H}' | \Psi_0 \rangle. \tag{2.20}$$

By adding both H' and $-H'$ to the right side of the equation, we can write:

$$\langle \Psi_0' | \hat{H} | \Psi_0' \rangle = \langle \Psi_0' | \hat{H}' | \Psi_0' \rangle + \langle \Psi_0' | (\hat{H} - \hat{H}') | \Psi_0' \rangle \tag{2.21}$$

$$= E_0' + \int d\mathbf{r} \rho(\mathbf{r})[V_{ext}(\mathbf{r}) - V_{ext}'(\mathbf{r})]. \tag{2.22}$$

Also, we can reverse the conditions and take $|\Psi_0\rangle$ as a trial wave function for the Hamiltonian $\hat{H}'$:

$$\langle \Psi_0 | \hat{H}' | \Psi_0 \rangle = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle + \langle \Psi_0 | (\hat{H}' - \hat{H}) | \Psi_0 \rangle \tag{2.23}$$

$$= E_0 + \int d\mathbf{r} \rho(\mathbf{r})[V_{ext}(\mathbf{r}) - V_{ext}'(\mathbf{r})]. \tag{2.24}$$

Re-writing the inequalities from Equations 2.19 and 2.20 in terms of the forms shown in Equations 2.22 and 2.24:

$$E_0' < E_0 + \int d\mathbf{r} \rho(\mathbf{r})[V_{ext}(\mathbf{r}) - V_{ext}'(\mathbf{r})] \tag{2.25}$$

$$E_0 < E_0' + \int d\mathbf{r} \rho(\mathbf{r})[V_{ext}'(\mathbf{r}) - V_{ext}(\mathbf{r})] \tag{2.26}$$

Then, by taking the summation of Equations 2.26 and 2.26, we find that

$$E_0 + E_0' < E_0 + E_0', \tag{2.27}$$

which is clearly a contradiction. Therefore, the external potential $V_{ext}(\mathbf{r})$ is uniquely determined by the ground state density $\rho(\mathbf{r})$. In addition, the number of electrons can be

determined by integrating the density over all space.

$$N = \int d\mathbf{r}\rho(\mathbf{r}).$$ (2.28)

The energy of the system can be written as

$$E_V[\rho(\mathbf{r})] = F[\rho(\mathbf{r})] + \int d\mathbf{r}V(\mathbf{r})\rho(\mathbf{r}),$$ (2.29)

where F is the electronic Hamiltonian from Equation 2.15.

2.3.4 Kohn-Sham equations

The variational problem for N electrons using the Hohenberg-Kohn density functional method can be written as,

$$\delta\left[F[\rho(\mathbf{r})] + \int d\mathbf{r}V_{ext}(\mathbf{r})\rho(\mathbf{r}) - \mu\left(\int d\mathbf{r}\rho(\mathbf{r}) - N\right)\right] = 0,$$ (2.30)

where μ is a Lagrange multiplier. First shown by Kohn and Sham, $F[\rho(\mathbf{r})]$ can be separated:

$$F[\rho(\mathbf{r})] = T_{ni}[\rho(\mathbf{r})] + \frac{1}{2}\int d\mathbf{r}d\mathbf{r'}\frac{\rho(\mathbf{r})\rho(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|} + E_{xc}[\rho(\mathbf{r})],$$ (2.31)

where T_{ni} is the classical kinetic energy from Equation 2.11. From Equations 2.30 and 2.31, we rearrange so that

$$\frac{\delta T_{ni}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})} + V_{KS}(\mathbf{r}) = \mu,$$ (2.32)

where the Kohn-Sham potential is given by

$$V_{KS}(\mathbf{r}) = \int d\mathbf{r}\frac{\rho(\mathbf{r})}{|\mathbf{r}-\mathbf{r'}|} + V_{xc}(\mathbf{r}) + V_{ext}(\mathbf{r}).$$ (2.33)

The exchange-correlation potential is related to the exchange-correlation functional by,

$$V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}.$$ (2.34)

It is important to note that Equation 2.32 would also be valid for a system of non-interacting particles experiencing an external potential $V_{KS}(\mathbf{r})$.

In order to find the ground state density, $\rho_0(\mathbf{r})$, which corresponds to the energy minimum of the electronic system, we use the simple one-electron Schrödinger equation

based on Equations 2.2 and 2.3. This can be rewritten as

$$\left[-\frac{1}{2}\nabla_i^2 + V_{KS}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}). \tag{2.35}$$

The ground state density can be found from the wave functions by this relation:

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N/2}\mid\psi_i(\mathbf{r})\mid^2, \tag{2.36}$$

where the factor of two is due to spin degeneracy from the assumption that the orbitals are singly-occupied.

2.3.4.1   Variational principle

We have shown in the above section that the external potential and ground state wave function are uniquely determined by the ground state density, and now we will show how to determine the ground state density. The variational principle dictates that every trial wave function except the unique ground state wave function will give an energy higher than the ground state energy:

$$\langle H\rangle = \frac{\langle\psi\mid H\mid\psi\rangle}{\langle\psi\mid\psi\rangle} \geq E_0 \tag{2.37}$$

where E_0 is the smallest eigenvalue of H. Expressed in terms of the electron density, $\rho(\mathbf{r})$, the calculated energy, $E_V[\rho(\mathbf{r})]$, is larger than the ground state energy, E_0, which is a minimum:

$$E_V[\rho(\mathbf{r})] \geq E_0. \tag{2.38}$$

This variational principle on the energy can be proven by using $\rho(\mathbf{r})$ which determines $V_{ext}(\mathbf{r})$ and ground state $|\Psi\rangle$. Using this state as a trial state for the external potential $V(\mathbf{r})$, we can write the total energy of the test system,

$$\langle\Psi|\hat{H}|\Psi\rangle = \langle\Psi|\hat{F}|\Psi\rangle + \langle\Psi|\hat{V}|\Psi\rangle \tag{2.39}$$

$$= F[\rho(\mathbf{r})] + \int d\mathbf{r}V(\mathbf{r})\rho(\mathbf{r}) \tag{2.40}$$

$$= E_V[\rho(\mathbf{r})] \geq E_0. \tag{2.41}$$

By minimizing the functional $E_V[\rho(\mathbf{r})]$, the energy will eventually approach but never meet the ground state energy, E_0.

2.3.5   Self-consistent procedure

Because the Kohn-Sham potential $V_{KS}(\mathbf{r})$ depends on the density $\rho(\mathbf{r})$, the equations must be solved self-consistently. First, a guess is made for density. The Schrödinger equation is solved (Equation 2.35) and a new set of orbitals $\{\psi_i(\mathbf{r})\}$ is determined. From these orbitals, a new density is found based on Equation 2.36. This new density then becomes the next guess for the Schrödinger equation, which gives new orbitals and another new density. This process is continued until the output and input densities are equivalent, indicating the ground state energy.

2.3.6   DFT implementation of exchange and correlation

So far, the exact nature of the exchange-correlation functional E_{xc} has not been discussed other than that it is the "quantum correction" to the classical kinetic and potential energies, and that it should include energetic corrections for fermionic spin exchange. We write E_{xc} as a function of ϵ_{xc}, the 'energy density'

$$E_{xc}[\rho(\mathbf{r})] \equiv \int d\mathbf{r}\, \epsilon_{xc}[\rho(\mathbf{r})]\rho(\mathbf{r}) \tag{2.42}$$

2.3.6.1   Local density approximation (LDA)

If ϵ_{xc} is computed based solely on the local position $\mathbf{r}$, then the method is called the Local Density Approximation (LDA). Found by Thomas-Fermi theory and Quantum Monte Carlo methods for a homogeneous electron gas [34], the exchange-correlation energy for LDA is written:

$$\epsilon_{xc}^{LDA}[\rho(\mathbf{r})] = \epsilon_x[\rho(\mathbf{r})] + \epsilon_c[\rho(\mathbf{r})] = -\frac{0.458}{r_s} - \frac{0.44}{r_s + 7.8}, \tag{2.43}$$

where

$$r_s = \left(\frac{3}{4\pi\rho(\mathbf{r})}\right)^{1/3}. \tag{2.44}$$

2.3.6.2   Generalized gradient approximation (GGA)

Because this electron-gas model does not regularly predict accurate chemical bonding properties, the next term in the expansion of the density is typically used. This is called the Generalized Gradient Approximation (GGA), in which

$$\epsilon_{xc}^{GGA}[\rho(\mathbf{r})] = \epsilon_{xc}^{LDA}[\rho(\mathbf{r})] + \Delta\epsilon_{xc}\left[\frac{|\,\nabla\rho(\mathbf{r})\,|}{\rho^{4/3}(\mathbf{r})}\right]. \tag{2.45}$$

GGA has shown remarkable flexibility for various chemical systems [35], especially covalent bond energies and distances, although known shortcomings are a poor description of van der Waals bonding [36] and incorrect unoccupied orbital energies leading to underestimation of band gaps for semi-conductors [37].

2.3.6.3   Hybrid DFT methods

Primarily, we have used B3LYP: the Becke three-parameter hybrid functional [38]. The exchange (x) and correlation (c) energy is written as E_{xc}^{B3LYP}, where

$$E_{xc}^{B3LYP} = (1-a)E_x^{LDA} + aE_x^{HF} + b\Delta E_x^{Becke} + E_c^{LDA} + c\Delta E_c^{LYP}, \tag{2.46}$$

and the coefficients were optimized to match extensive molecular data sets (a=0.20, b=0.72, and c=0.81) [38]. The first term in the hybrid method is E_x^{LDA}, which is the LDA exchange term from Equation 2.43. E_x^{HF} is Hartree-Fock exchange integral, which is exact and is given in Equation 2.12. Becke's B88 exchange term [39] is based on empirical results, and is written as,

$$E_x^{Becke}[\rho(\mathbf{r})] = -\beta \int d\mathbf{r}\rho(\mathbf{r})^{4/3} \frac{\alpha^2}{(1+6\beta\sinh^{-1}\alpha)} \tag{2.47}$$

where

$$\alpha = \frac{\mid \bigtriangledown\rho(\mathbf{r}) \mid}{\rho(\mathbf{r})^{4/3}}.$$

Found by matching molecular data sets, β was found to be 0.0042 Hartree. Correlation functionals are from the LDA [40] and from Lee, Yang, and Parr (LYP) [41, 42], the latter based on an empirically determined model of the correlation energy of electrons in a helium atom.

Implemented with Gaussian code [43], this hybrid gradient-corrected method is considered one of the most accurate exchange-correlation functionals and has been used with great success in other biological systems [44, 45]. Calculations with post-Hartree-Fock Møller-Plesset perturbation theory (MP2) [46, 47, 48, 49] were conducted to test the accuracy of the B3LYP method for this system, and the energy barrier calculations were consistent.

2.3.7 Basis sets

Any numerical basis set may be used, although for computational efficiency only a several are implemented. A primitive Gaussian-type orbital (GTO) in atom-centered Cartesian coordinates is

$$\phi(x, y, z; \alpha, i, j, k) = \left(\frac{2\alpha}{\pi}\right)^{3/4} \left[\frac{(8\alpha)^{i+j+k} i! j! k!}{(2i)!(2j)!(2k)!}\right]^{1/2} x^i y^j z^k e^{-\alpha(x^2+y^2+z^2)} \qquad (2.48)$$

where α controls the width of the GTO, and i, j, and k are non-negative integers that define the angular momentum of the orbital. For example, for an s-type GTO (orbital angular momentum $l = 0$), the integers are $i = j = k = 0$, and ϕ is spherically symmetric. For a p-type GTO (orbital angular momentum $l = 1$), there are three possibilities for i, j, and k that lead to the Cartesian prefactor of either x, y, z. For d-type GTOs (orbital angular momentum $l = 2$), there are six possibilities for the index values (i, j, k). Specifically, the Cartesian prefactors are $x^2, y^2, z^2, xy, xz, yz$. All of these Cartesian prefactors are multiplied by the Gaussian term in order to obtain proper atomic orbital shape.

GTOs are easily integrated by computational schemes, but they do not accurately enough resemble atomic orbitals (AOs). Primitive GTOs are smooth and differential at the nucleus ($r = 0$). But for a hydrogenic system, the actual atomic orbital (AO) will have a non-continuous derivate at the $r = 0$, which would be better described r^{-1}, a less integrable function (called Slater-type orbital, or STO). In order to use a basis function with the proper radial shape, and one that when squared is easily integrable *via* numerical methods, we have used contracted Gaussian functions, which are the linear combination of primitive GTOs designed to resemble an STO. A general contracted GTO may be written as

$$\varphi(x, y, z; \{\alpha\}, i, j, k) = \sum_{a=1}^{M} c_a \phi(x, y, z; \alpha_a, i, j, k) \qquad (2.49)$$

where M is the number of GTOs used in the linear combination. The coefficients c_a are chosen for normalization as well as to optimize the basis function shape, which are then called a linear combination of atomic orbitals (LCAO).

We have used the double-ζ basis set, 6-31G(d,p), for geometry optimizations during initial reaction path sampling [50], where the '6' represents six GTOs for core electrons and the '31' represents split GTOs for valence electrons: specifically three and one. Split-valence basis sets allow for a more accurate description of chemical bonding due to increased flexibility to fit valence electrons into molecular orbitals, and are the norm

when using a Gaussian-type basis set. The '(d,p)' indicates that we are using polarization functions that allow for a shift in the wave function away from the atomic center. We have also used the triple-ζ basis set, 6-311++g(d,p), for calculations of the local minima and transition states found with the first basis set [51]. Diffuse functions for long range interactions are represented with a '+', and are especially important for anions. Basis sets of similar size are typically used for systems with similar number of electrons, and our test calculations as well as the work of others have shown these basis sets to be sufficient for similar atom types [44, 45].

2.4   Calculation methodology

In addition to highly efficient computational codes available and tuned for parallel processing, we have written various useful computer codes for data analysis and manipulation, especially for trajectories of large QM/MM systems that contain thousands of atoms which would be extremely cumbersome otherwise. The electronic structure calculations are performed with robust computer code widely used in the field, and each method is described below.

2.4.1   Gas phase

Calculations *in vacuo* are the simplest way to understand the energies associated with a reaction mechanism, since they treat the quantum mechanical active site as an isolated molecule or group. Gas phase calculations capture the amount of energy necessary to transverse the energy barrier, which is typically then lowered by electrostatic interactions with the remaining protein and/or solvent. By studying mechanisms in the gas phase, we can separate the energy related to a particular bond breaking/forming from other factors such as structural rearrangement and polarization effects. One drawback of using a small model system with gas phase calculations is that the protein backbone is either rigid with the starting configuration, or completely relaxed and prone to unlikely rearrangement in the protein context. Despite these limitation, reaction profiles calculated with gas phase calculations are an important part in describing the energies of a reaction.

2.4.2   Semi-empirical PM3

Semi-empirical PM3 [52] calculations were performed in order to obtain a two-dimensional potential energy surface in the xy plane for two constrained reaction coor-

dinates. Geometry optimizations with PM3 are computationally efficient and sufficiently accurate for the present purposes. Deficiencies of PM3 with respect to chemical structures, such as the "flattening" of small and medium sized rings [53], nitrogen atoms with a lone pair having pyramidal geometries [54], and inaccurate hydrogen bonding distances are well known. Nevertheless, PM3 is useful for efficiently scanning many geometries and locating general locations of transition states, as is done here.

2.4.3   Implicit solvent

One method for approximating the environmental electrostatic effect is to use an implicit solvent. In this scheme, the active site is polarized by the dielectric medium which is itself polarizable. The Polarizable Continuum Model (PCM) [55] was used to simulate solvent effects in the detailed calculations. The numerical Integral Equation Formalism [56] (IEFPCM) was used because it allows for interlocking atomic spheres to represent the extent of the system in solution, which is important for protons that are in between atoms during a chemical reaction and at or around the energy barrier.

Non-dimensional dielectric constants are defined by $\epsilon_r = \epsilon_s/\epsilon_0$, where ϵ_0 is the vacuum permittivity and ϵ_s is the static dielectric constant for the dielectric. For the gas phase, $\epsilon_r = 1$. For water, $\epsilon_r = 78.39$. Geometry optimizations were performed in implicit solvent and results are compared with gas phase calculations.

2.4.4   Multiscale modeling

2.4.4.1   Combined quantum and classical mechanics

The Quantum Mechanics/Molecular Mechanics (QM/MM) layering method is used, and involves treating the protein active site and critical solvent molecules with first principles methods while treating the remaining full-protein system with classical force fields [57, 58]. Similar multiscale methods have been used with good success [59, 60, 61, 62, 63]. The classical system is periodic and is truncated to include the protein (intein and exteins) as well as all interior waters and certain exterior water molecules that are within a range of 7.0 Å to the protein surface. All atoms are relaxed, and each calculation includes at least 6000 atoms, roughly 2350 of which belong to the protein. The full-protein plus solvent system, termed the real system, is treated only with the classical MM method. Within the real system, the active site model system is partitioned and is treated independently by QM and MM methods. Dangling bonds that are introduced by partitioning

the model system are passivated with hydrogen atoms. With normal QM/MM energy calculations and geometry optimizations, protein and solution outside the model system is typically only included as a mechanical perturbation. For this reason, it is critical that the model system should include protein segments and solution molecules that are interacting electrostatically. The combined Hamiltonian may be written:

$$E_{ONIOM}^{QM:MM} = E_{Model}^{QM} - E_{Model}^{MM} + E_{Real}^{MM}. \tag{2.50}$$

2.4.4.2 Charge embedding

In addition to the mechanical perturbation on the QM Hamiltonian, the electrostatic contribution from the partial charges of the MM region can be included as a perturbation on the QM Hamiltonian. For this scheme the partial charges are those used in the MM calculation and are scaled by the default manner where atoms bonded to the inner-most four layers and atoms outside that threshold are not included [58]. Typically, we report E_{Model}^{QM}, which represents the QM active site energy. The other energy terms, including the combined $E_{ONIOM}^{QM:MM}$ involves classical parameters that have no relevance to the energies of bond forming and breaking at transition states.

2.4.5 Geometry minimization

Due to the complexity of biomolecular reactions, a rigorous multidimensional search over local conformational space is essentially required although not computationally feasible for large systems [64]. Due to the time expense for each calculation, we have used the constant minimization procedure. For intermediate states along the reaction path, one coordinate is constrained while the remaining system is relaxed. The constrained internal coordinate, called the Asn cyclization distance, was the atomic distance between the Asn side chain N atom and the carbonyl C of Asn on the scissile peptide bond. In calculations with a hydronium ion (H_3O^+), the three O-H bond distances were often constrained to 0.98 Å to avoid spontaneous proton donation observed otherwise.

2.4.6 Free energy calculation

Thermal and entropic contributions calculated with a harmonic approximation for the optimized geometries at the B3LYP/6-31G(d,p) level were combined with the electronic energy to obtain free-energy profiles in the gas phase and in the implicit solvent. Zero point energies were found to differ by between 0.04 and 1.33 kcal/mol, which are

within the expected error for the calculation. The approximate entropic components of the free energy include contributions from translational, electronic, rotational, and vibrational degrees of freedom [65, 66] and were obtained from frequency calculations at room temperature. Thermal corrections do not include imaginary frequencies of vibrational modes for transition states.

CHAPTER 3
C-TERMINAL CLEAVAGE

3.1 Introduction: C-terminal cleavage

3.1.1 Experimental motivation

C-terminal cleavage involves the isolated excision of the C-extein from the intein substrate. In the experimental study by Wood *et al.* on the *Mycobacterium tuberculosis* (*Mtu*) recA cleavage mutant mini-intein (ΔI-CM), a decrease in solution pH from 7.5 to 6.0 was found to lead to a significant increase in the rate of C-terminal cleavage [19, 24]. The higher C-terminal cleavage activity at lower pH for this intein as well as for the *Ssp* DnaB intein are, however, inconsistent with the currently available details of the mechanism proposed for intein cleavage within the context of splicing. For example, Ding *et al.* have suggested that for the reaction of the *Ssp* dnaB mini-intein, C2-His12 (F-block) acts as a base, deprotonating the nitrogen of the C-terminal C1-Asn8 side chain *via* a vicinal water molecule. Experiments on short peptides in solution also show an increased tendency of Asn to be ionized and to cyclize (leading to succinimide formation) over the pH range of 7.4-13.8 [67, 68, 69, 70]. For the side chain of His to act as a base and accept protons, one of its two imidazole side chain nitrogen atoms must be deprotonated on average. At lower pH, especially below the pK_a, His is more likely to be present in the doubly protonated and positively charged state, diminishing its ability to accept protons. Thus, the mechanistic details underlying the increased C-terminal cleavage activity observed in experiments of Wood *et al.* [19, 24] are expected to be different from those proposed by Ding *et al.* [71].

3.1.2 Reaction mechanism

We have studied the cleavage mechanism of the *Mtu* recA intein using information from available intein crystal structures and mutagenesis experiments. Part of the intein structure close to the C-terminal reaction site was considered in both gas phase as well as implicit solvent model calculations.

Based on a combination of semi-empirical and first principles computational analyses, and with gas phase, implicit solvent, and QM/MM calculations, we proposed new details of the mechanism of Asn cyclization catalyzed by the cleavage mutant of the *Mtu* recA mini-intein that account for an increase in activity of C-terminal cleavage at low pH
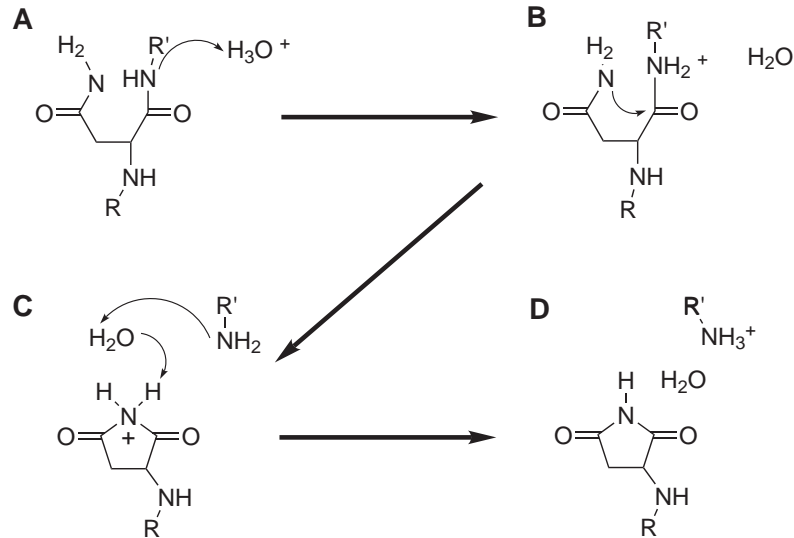
Figure 3.1: **The proposed N-protonation reaction scheme for Asn cyclization to succinimide; R and R' represent the intein and C-extein, respectively.**

[28]. Figure 3.1 shows important reaction steps. This mechanism involves the protonation of the nitrogen of the scissile peptide bond by a vicinal hydronium ion (H_3O^+). This leads to stretching of that peptide bond due to the loss of π-bond resonance and consequent increase in carbon electrophilicity [72]. Asn cyclization and subsequent succinimide formation occur, resulting in peptide bond cleavage. Our results are consistent with the experimental observations of Wood *et al.* that indicate a simple proton-catalyzed reaction [19, 24]. The frozen internal coordinate, called the Asn cyclization distance, is the atomic distance between the Asn side chain N atom and the carbonyl C of Asn on the scissile peptide bond (shown with an arrow in Figure 3.1B).

Given current computational resources, detailed quantum mechanical calculations are limited in system size. In the present context, where the mechanism of enzymatic catalysis is of interest, one may choose on the order of 10^2 atoms and perform only a few calculations, or choose on the order of 10^1 atoms that is relevant part of the system, and explore a variety of possible reaction pathways. Inclusion of a larger number of atoms not only adds to the computational expense, but could allow conformational rearrangements that are inconsistent with the protein structural context being studied. As a compromise, and based on available crystal structure knowledge, our initial system includes 25 atoms: the C-terminal C1-Asn8 side chain, the backbone atoms of the penultimate C1-His7 and of the C-terminal Cys+1 (the dangling bonds are passivated with hydrogen atoms), and
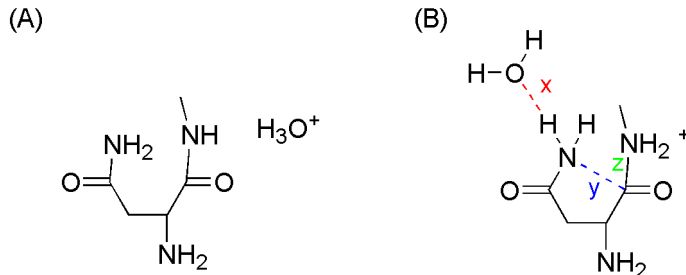
Figure 3.2: PM3 reaction coordinate. The system before N-protonation is shown (A), and the reaction coordinate for the energy profile shown in Figure 3.3 is highlighted (B).

one explicit water molecule. Additional calculations with larger systems containing 50+ atoms, and a full protein quantum mechanics/molecular mechanics (QM/MM) treatment indicated energetic results will be described in Chapter 4.

3.2 Semi-empirical analysis

3.2.1 Description of model system

The PM3 Hamiltonian was used for scanning the potential energy surface in two dimensions. Given that the doubly protonated state of the amide nitrogen is a likely starting point of the C-terminal cleavage reaction at low pH, we performed geometry optimizations using the PM3 method for 428 independent points in the two dimensional space based on reaction coordinates x and y (shown in Figure 3.2). Coordinate x represents the distance between the oxygen atom of the water molecule and the hydrogen atom of the Asn side chain, referred to as the Asn ionization distance. Coordinate y, the Asn cyclization distance, is the separation between the Asn side chain nitrogen and the peptide carbonyl carbon.

3.2.2 Reaction coordinate space

The space of reaction coordinates for the semi-empirical calculation is shown in Figure 3.2 and corresponding energies for the two-dimensional energy scan are shown in Figure 3.3. The Asn ionization distance, x, ranges from 0.9 to 3.0 Å ($x=2.0$ Å corresponds to a typical hydrogen bond, whereas $x=1.0$ Å indicates the deprotonated Asn side chain and a re-formed hydronium ion). The Asn cyclization distance, y, ranges from 1.4 to 3.5 Å ($y=3.5$ Å is the relaxed distance in the initial state, whereas $y=1.5$ Å indicates fully cyclized Asn). The initial state ($x=3.0$, $y=3.5$ Å) located on the top-right of the plot is
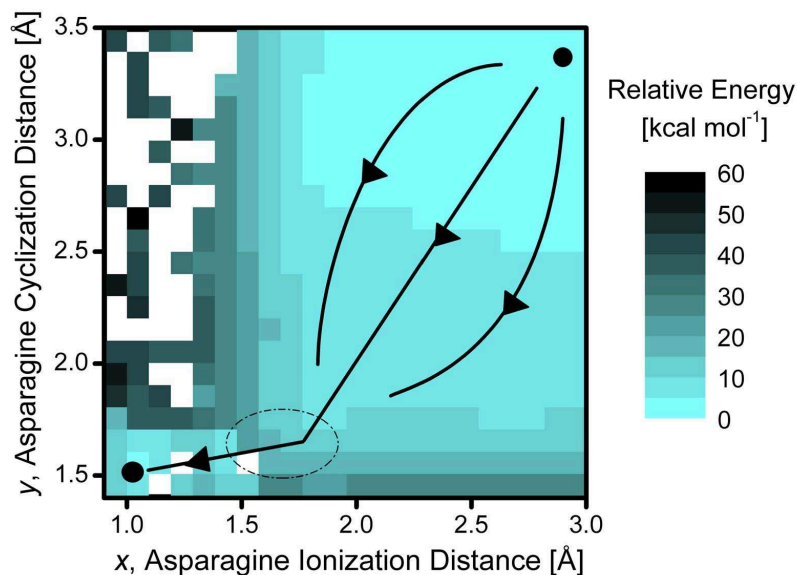
Figure 3.3: Semi-empirical (PM3) energy surface shown for the reaction coordinates shown in Figure 3.2. Filled circles indicate reactant and product state, and the arrows indicate proposed reaction path. The dotted ellipse indicates the general location of the energy barrier.

chosen to be zero for the relative reaction energy. The final products state is located in the bottom left corner of the graph (x=1.0, y=1.5 Å) and corresponds to cyclized Asn (succinimide), a re-formed hydronium ion, and a cleaved peptide bond.

3.2.3 Reaction path

The path marked by arrows on Figure 3.3 indicates the likely path followed by the C-terminal cleavage reaction. Along that path, y is reduced significantly first, and is then followed by a reduction in the value of x. The reduction in y may happen along combinations of paths shown in Figure 3.3 because that region of the energy landscape is relatively featureless. In any case, cyclization of Asn appears to be almost complete before the ionization of the side chain nitrogen takes place. The barrier region is located near $x = 1.6$, $y = 1.6$ Å indicated by the ellipse in Figure 3.3 and has energy of about 25 kcal/mol higher than the reference state. Figure 3.3 shows that alternate paths, *i.e.*, Asn ionizes before its cyclization, and are highly unlikely as they sample regions of considerably high energies.

We note that PM3 calculations for geometry optimizations do not converge for certain choices of x and y (open squares in Figure 3.3). Most of these points are neighbored by points of higher energies, and therefore, should not affect the general conclusions drawn above.

3.3   First principles reaction study

3.3.1   Gas phase molecular system

As will be apparent in the discussion below, protonation of the backbone nitrogen of the scissile peptide bond is a necessary first step in the reaction at low pH. To this end, Figure 3.4 considers three scenarios corresponding to i), a neutral peptide case in which neither of the atoms are protonated, ii), O-protonation of the carbonyl oxygen, and iii), the N-protonation of the peptide nitrogen [72] Specifically, we study consequences of these three scenarios on the system energy and the equilibrium length of the scissile peptide bond. In a neutral peptide (scenario i above), the relaxed peptide bond length is 1.35 Å. As Asn cyclization proceeds (*i.e.*, as y is reduced), the system energy increases significantly (Figure 3.4, bottom graph). Correspondingly, there is only a slight increase in the peptide bond length, indicating that it remains essentially intact.

When the carbonyl oxygen atom is protonated (scenario ii), the relaxed peptide bond length in fact decreases to 1.32 Å, as expected from the increased π-conjugation (or the double bond character of the bond) between C and N. Asn cyclization energy (Figure 3.4, top plot) in this case is lower than that for the neutral peptide case; however, the peptide bond is significantly more stable and remains essentially intact as Asn cyclization proceeds.

In contrast, when the peptide nitrogen atom is protonated (scenario iii), the relaxed peptide bond length increases to 1.51 Å, indicating the weakening of that bond. As Asn cyclization proceeds, that distance increases further and leads to breaking of that bond, resulting from the fact that a doubly protonated nitrogen makes a good leaving group (Figure 3.4, bottom plot). The cyclization energy (Figure 3.4, top plot) in this case is lower than that for the neutral peptide case and similar to that for oxygen protonation, which does not lead to peptide bond cleavage (see above). Collectively, these preliminary calculations indicate that protonation of the amide nitrogen is an important first step for C-terminal cleavage in low pH environments.

The protonation of the nitrogen of the scissile peptide bond proposed above makes
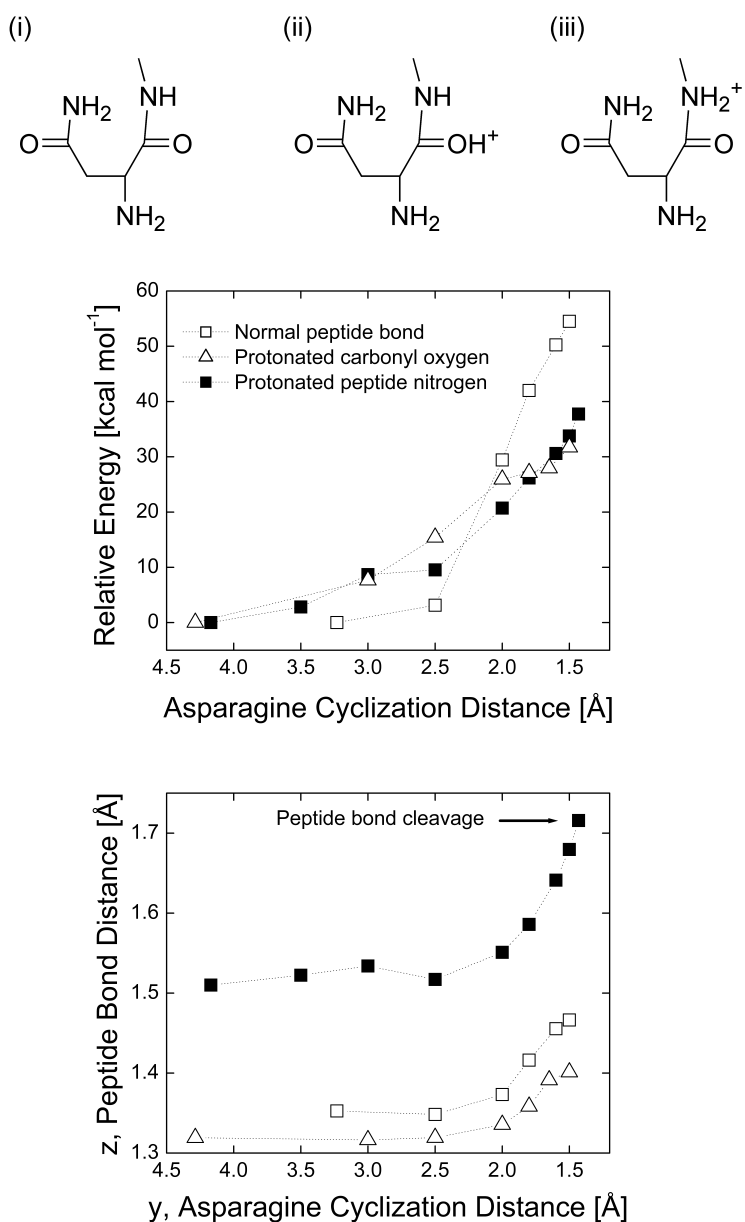
Figure 3.4: Schematic of the three scenarios considered for Asn cyclization: i), normal peptide, ii), O-protonation, and iii), N-protonation. Asn cyclization energies (top graph) and peptide bond stretching (bottom graph) versus the Asn cyclization distance, y, for the neutral peptide bond system ($\square$), for the system where the carbonyl oxygen is protonated ($\triangle$), and where the peptide nitrogen atom is protonated ($\blacksquare$). Gas phase structures are optimized at the B3LYP/6-31G(d,p) level.

that nitrogen atom transiently doubly protonated. In the broader context of enzymatic catalysis, this proposal is not new. Indeed, in the hydrolysis of a peptide bond by serine proteases, the nitrogen of the scissile peptide bond accepts a proton from the His of the catalytic triad [73]. In their study of enzymatic reaction catalyzed by the HIV-1 protease, Trylska *et al.* found that protonation of the amide nitrogen was essential for peptide bond cleavage [74]. Similarly, the protonation of the amide nitrogen was found to be the essential step in the hydrolysis of a formamide molecule, which was used as a computational model for peptide bond hydrolysis [75].

The observation made above that the Asn cyclization proceeds before the ionization of its side chain is supported independently by high-level quantum calculations at the B3LYP/6-31G(d,p) level. Specifically, we followed the ionization of the Asn by gradually transferring the proton from the side-chain nitrogen to the vicinal water molecule (Figure 3.5, top panel). These high-level calculations also included effects of the dielectric constant of the local environment, which was assumed to be equal to 1 in the (gas phase) PM3 calculations.

3.3.2   Energetic results

The graph in the bottom panel of Figure 3.5 shows that both the electronic energy (E) and Gibbs free energy (G) for the ionization of the Asn side chain are rather high, equal to ∼31 kcal/mol and ∼35 kcal/mol, respectively, even in the highly polar medium such as water. The intein active site is expected to have a dielectric constant lower than that of water, and therefore, the relative energy of ionization will be even higher. We note that the Asn cyclization distance, y, is relaxed in these calculations and does not reduce significantly. Thus, Asn cyclization will require additional energy. In contrast, as shown later, the side-chain ionization is almost spontaneous once the Asn side chain has undergone cyclization and formed a succinimide, consistent with the experimental observations of enhanced cleavage at low pH by Wood *et al.* [19, 24].

3.3.3   Mechanistic details

Collectively, the above calculations allow us to propose a somewhat detailed C-terminal cleavage reaction mechanism at low pH, in which six states shown in Figure 3.6 are particularly important. Figure 3.6A shows the hydronium ion in the context of the relevant part of our intein system. The second state involves the donation of a proton
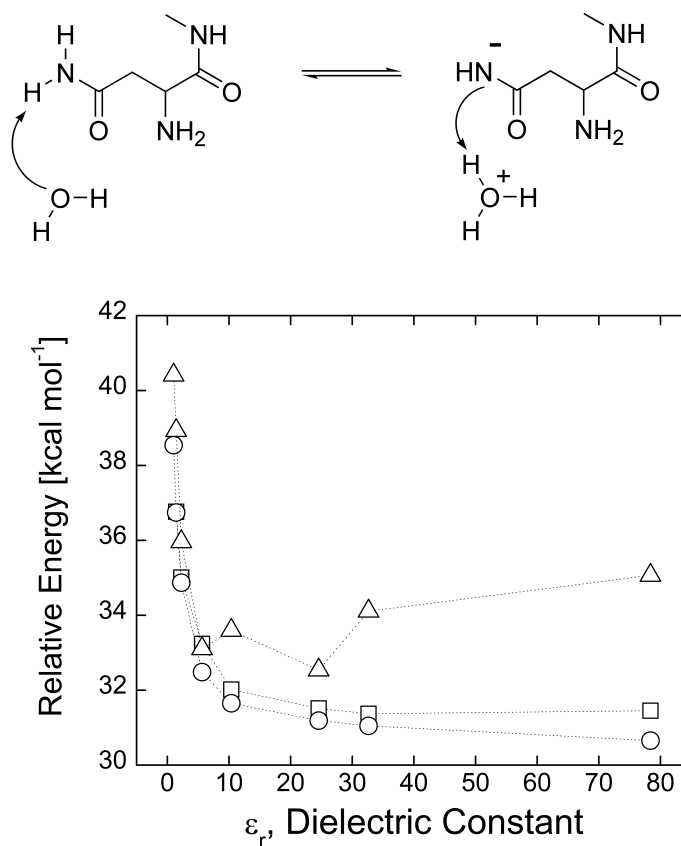
Figure 3.5: Energy of ionization of Asn side-chain nitrogen (top panel) calculated in solvents of different dielectric constants: vacuum (ϵ_r =1.000), argon (ϵ_r =1.43), benzene (ϵ_r =2.247), chlorobenzene (ϵ_r =5.621), dichloroethane (ϵ_r =10.36), ethanol (ϵ_r =24.55), methanol (ϵ_r =32.63), and water (ϵ_r =78.39). Each implicit solvent has unique parameters such as radius and density. Note that the peptide nitrogen has only one proton in this case. Gas phase coordinates, energy in solvent ($\bigcirc$); solvent optimized coordinates, energy ($\square$); and Gibbs free energy ($\triangle$).
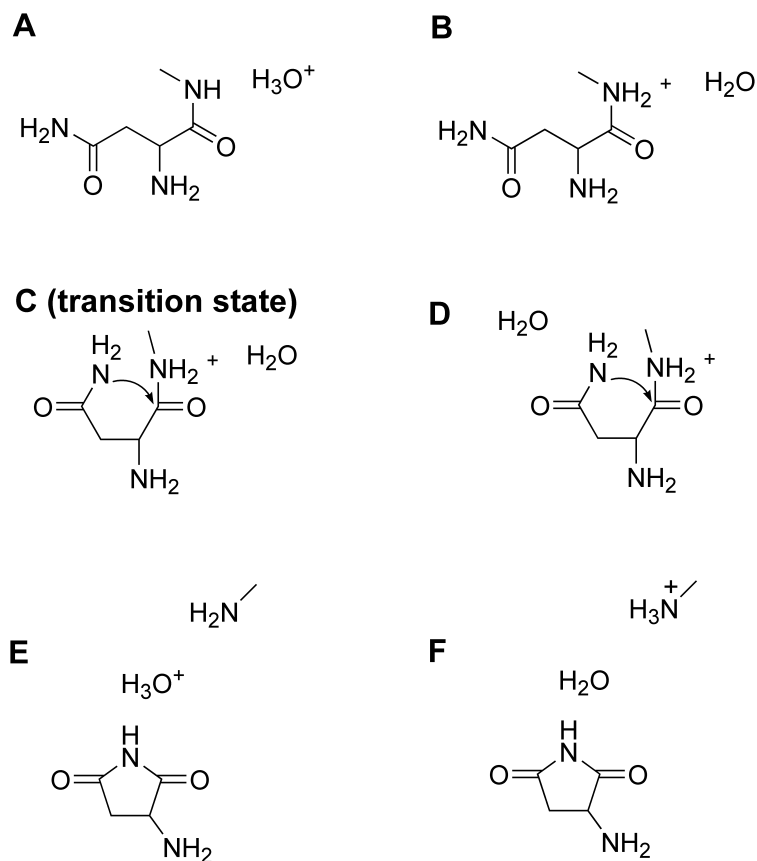
Figure 3.6: **Important states in the proposed hydronium ion catalyzed Asn cyclization and peptide bond cleavage mechanism (see text for details).**

by the hydronium ion to the peptide nitrogen, resulting in water and N-protonated state (Figure 3.6B). Asn cyclization is shown in Figure 3.6, C and D, where the Asn side chain still has two protons. The explicit water molecule is adjacent to the peptide nitrogen in one case (Figure 3.6 C) and moves to accept a proton from the forming succinimide in another (Figure 3.6D). The formed succinimide with the proton passed back to the hydronium ion is shown in Figure 3.6E, whereas the final product is shown in Figure 3.6F. Water is re-formed and the extein segment leaves an a terminal -NH_3^+ group.

The vicinal water molecule plays an important role in this mechanism and is used both as an acid (state A/B) and a base (state D/E). Indeed, succinimide with NH_2 is highly acidic due to the resonance effect of amide bonds on either side of that nitrogen. As a result, the nitrogen readily gives a proton to a nearby water molecule (state F).

The energies corresponding to the various states (from A to F) presented in Figure 3.6 are shown in Figure 3.7. State Z in Figure 3.7 is shown as a reference. The transition
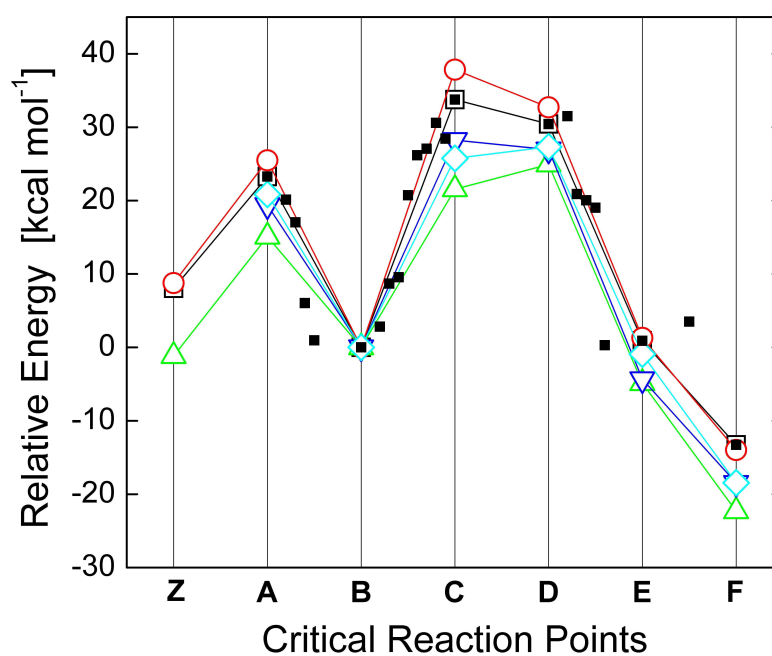
Figure 3.7: Electronic energies and approximate Gibbs free energies for states A through F in Figure 3.6 (energies are relative to state B). State **Z** is shown as a reference and corresponds to a positively charged His and a water molecule (**Z**). Figure key: vacuum optimized coordinates energy ($\square$) and Gibbs free energy ($\bigcirc$); water optimized coordinates, energy ($\triangle$); and Gibbs free energy ($\triangledown$). Single point energies calculated with B3LYP/6-311++g(2d,p) using the coordinates optimized in implicit water without the diffuse functions ($\diamond$). Energies of intermediate structures between the states in Figure 3.6 optimized in the gas phase ($\blacksquare$) are also shown.

from Z to A corresponds to a positively charged His and a water molecule (Z) forming a neutral His and a hydronium ion (state A). The energy barrier from B to C/D corresponds to the Asn cyclization, where the second proton is still attached to the Asn side chain. Indeed the normal mode corresponding to the single imaginary frequency in state C was in the direction of bond formation, as expected. For this system, the energetics suggest that the nitrogen will give its second proton to water to reform the hydronium ion (Figure 3.6E). As discussed above, in the absence of protonation of the peptide nitrogen (*e.g.*, in the case of a neutral peptide), the Asn cyclization has a higher barrier.

Figure 3.6 also highlights the effects of taking into account the dielectric constant of the environment of the active site on intein. As expected, the energy barrier ($\sim$33 kcal/mol) in the gas phase is reduced to $\sim$25 kcal/mol in implicit solvent having a high dielectric constant. When tested with MP2, the energy barrier was found to be 29.4 kcal/mol at the MP2/aug-cc-pVDZ level. Since the crystal structures of active inteins do not include exteins, the conformation of the exteins and N- and C-terminal active sites is unknown. Hence, the intein plus extein system used in the QM/MM calculation will require additional verification to be comparable to precursor inteins used in experiment. The actual dielectric constant of the protein interior is between that of bulk water and vacuum, and hence, our calculations suggest that the energy barrier for C-terminal cleavage lies between 25 and 33 kcal/mol. The experimental value is $\sim$21 kcal/mol at pH 6.0 [19, 24].

We note that these initial calculations have several limitations. In the actual intein system, the overall protein structure (including both the intein and exteins) that surrounds the active site provides a significantly greater structural as well as chemical context for the reaction to occur. Also, there will likely be more than one water molecule in the vicinity of the active site that could mitigate the C-terminal cleavage reaction. Our system, in contrast, is significantly smaller due to computational limitations. In addition, our calculations are by necessity static in nature, and ignore the conformational and water exchange dynamics that are important in enzymatic catalysis [76]. These types of concerns are shared by most (if not all) quantum calculations of enzymatic reactions. Nevertheless, these initial calculations provide a plausible mechanism for C-terminal cleavage that are tested in the next chapter by using larger computational systems and application of better multi-scale methods.

It should be noted that the possibility of protonation by His is not excluded by the

mechanism proposed here. For example, positively charged His could donate a proton to the peptide nitrogen via a water molecule; the reaction can then follow steps similar to the ones outlined here.

CHAPTER 4
C-TERMINAL CLEAVAGE - EFFECT OF +1 EXTEIN RESIDUE

The previous chapter described preliminary results in determining the low pH reaction mechanism for intein C-terminal cleavage. Here, the C-terminal cleavage reaction mechanism is further discussed in the context of a larger computational system (at least 45 atoms) as well as within the framework of the coupled Hamiltonian approach of quantum mechanics/molecular mechanics (QM/MM).

4.1 Non-essential mutation

Once splicing was inhibited, the downstream Cys residue (which was the first amino acid of the C-terminal extein or C-extein) was found to be functionally unnecessary for the C-terminal cleavage mechanism. Interestingly, Wood *et al.* observed that this amino acid regulated the reaction rate but did not alter the mechanism [24]. Furthermore, since the CM was found to be exceedingly reactive at low pH values, Wood *et al.* [24] utilized Met, which was the native N-terminus of the protein that formed the C-extein sequence, to decrease the reaction rate by an order of magnitude. In this experiment, three proteins of various sizes were contrasted with only the Cys/Met C-extein mutation: Thymidylate synthase (31.5 kDa), Hfq Protein (18 kDa), and rh aFGF (14 kDa). For these proteins, the Cys to Met mutation resulted in a decrease of the reaction rate by a factor of 12.0, 5.0, and 7.8, respectively [24, 26]. Figure 1.3 shows a schematic of the intein precursor and products based on these results [10, 27], although the exact mechanisms that govern the splicing and cleavage reactions are not understood at the atomic level. In particular, the effect of the single amino acid mutation at C+1, flanking the conserved C1: His7-Asn8 dipeptide at the intein terminus, on the reaction rate is not understood.

In order to obtain an atomic-level understanding of the effect of mutation on the reaction barrier, detailed quantum mechanical calculations on the intein C-terminal cleavage reaction have been carried out [28]. Simulations were based on both full quantum mechanical molecular analysis as well as a hybrid quantum mechanics and molecular mechanics (QM/MM) approach where the entire protein and solvent are treated classically

with parameterized force fields in a molecular mechanics (MM) calculation as shown in Figure 4.1(a). The 53 atom C-terminal catalytic site (C1-block: His–Asn–Cys, or His–Asn–Xxx, where Xxx is an alternate amino acid) was treated with quantum mechanics (QM) and is shown in Figure 4.1(b).

The computational energy barrier was smaller for the C-terminal sequence His–Asn–Cys than for that of the His–Asn–Met mutant, consistent with experimental observations [24, 26]. The difference in energy barrier between Cys/Met residues was due to the difference in electron affinity of the amino acids. In addition to Cys and Met, several other amino acids at the first C-extein position (C+1) were studied here. The energy barrier for C-terminal cleavage, calculated with a larger model system, is confirmed to match experiment.

4.1.1 Classical protein system

Starting with the intein crystal structure for the *Mtu* recA intein, ($\Delta\Delta$Ihh-CM, PDB code 2IN8) [11], a product protein without exteins, N- and C-terminal exteins were computationally added and then equilibrated with classical molecular dynamics (MD) simulations. The N-extein sequence consisted of Ace-Val-Val-Lys-Asn-Lys and the C-extein sequence consisted of Cys-Ser-Pro-Pro-Phe-Nme, both based on the native extein sequences [77]. Ace and Nme were capping residues for the N and C-terminal exteins, respectively. AMBER force field parameters [78] were implemented with GROMACS code [79]. MD simulations were carried out for 4 ns (0.5 ns equilibration, 3.5 ns production run) with temperature T = 298 K, pressure = 1 bar, and number of water molecules = 9548 for Cys and 9549 for Met systems.

4.1.2 Tripeptide subsystem

4.1.2.1 Description of model system

The tripeptide active site system (His–Asn–Cys) is highlighted in the view of the full intein crystal structure in Figure 4.1(b). Gas phase calculations were used to study the effect of site-directed mutagenesis (see Figure 4.2). Intein crystal structures usually include a hydrogen bond between the N^δ–H of the (penultimate) His side chain and the carbonyl O of Asn, the final amino acid of the intein [12, 13, 14, 15, 16]. Although the penultimate intein His residue has been previously assumed to be the proton donor for C-terminal cleavage reaction in the context of splicing [71], further inspection revealed that this was
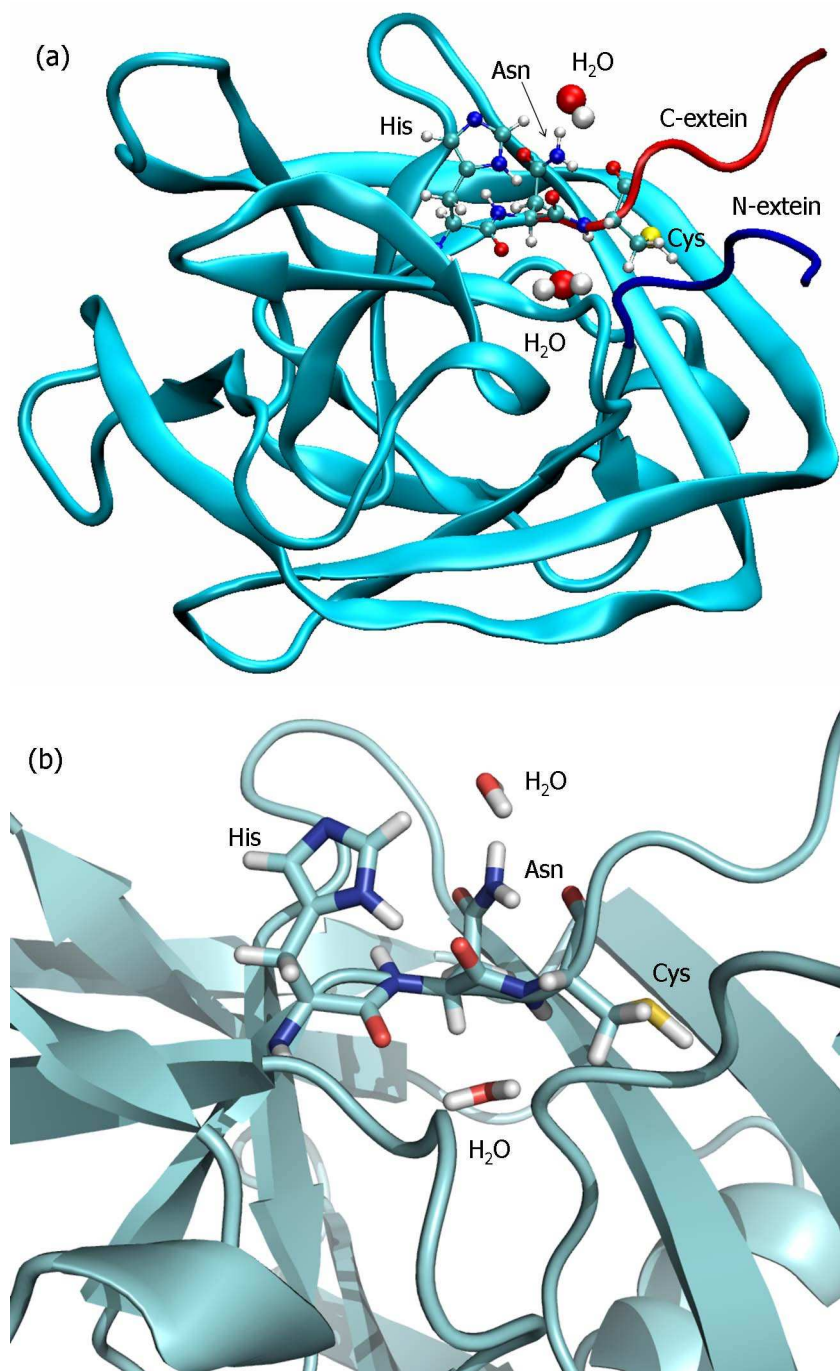
Figure 4.1: The intein cleavage mutant (CM) crystal structure (PDB code 2IN8) with computationally added exteins (a). The C-terminal catalytic site (His–Asn–Cys + two water molecules) is highlighted (b).
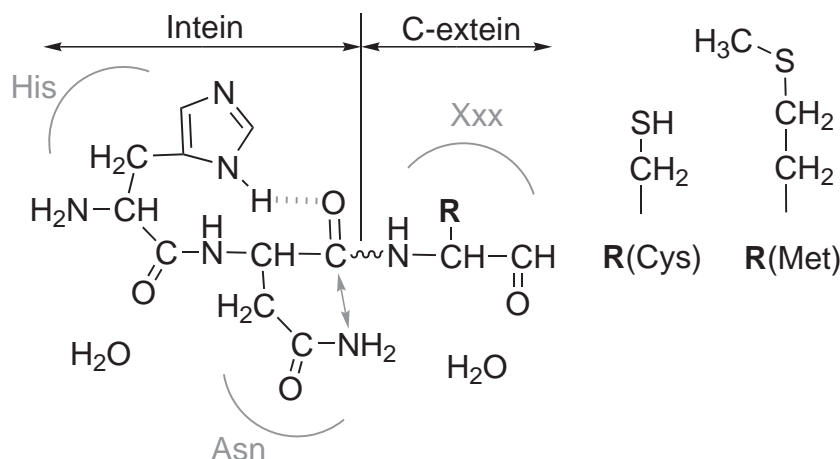
Figure 4.2: The C1-block His–Asn–Xxx active site is shown. The highly conserved H-bond is shown with a dotted line, the cyclization coordinate of Asn is shown with an arrow, and the scissile peptide bond is shown with a wavy line. Side chains for Cys and Met are shown, although Ala, Val, Thr and Ser were also considered.

not the case for pH dependent C-terminal cleavage. For a simple proton-catalyzed reaction, there is an inverse linear rate dependence on the pH, which was observed experimentally for the C-terminal cleavage reaction [24]. Since the ability of His to act as an acid is based on its local pK_a value, the expected pH-rate curve should be non-linear, specifically sigmoidal in shape, which is in contrast to the linearity observed experimentally.

The proposed N-protonation mechanism begins with the protonation of the peptide N by a hydronium ion (H_3O^+, see Figure 3.1). This in turn causes the scissile peptide bond to elongate, and hence reduces the energy necessary for peptide bond cleavage after Asn cyclization. After Asn cyclization and aminosuccinimide formation, the extra proton passes to the cleaved C-extein N-terminus (-NH_2), which is excised and leaves with a positive charge (-NH_3^+, see Figure 3.1D). Although O-protonation was more energetically favorable for a generic or average peptide that was fully solvent exposed, in the case of the intein C-terminal active site, the carbonyl O was strongly hydrogen bonded to the N^δ–H of His and was also pointed inward, toward the core of the protein and away from the main body of solvent. The Asn cyclization reaction after O-protonation instead of N-protonation has been shown to require more energy and does not lead to cleavage of the peptide bond [28].

Prior to the QM/MM full protein study, the His–Asn–Cys tripeptide system (Figure

4.2) was studied with an isolated gas phase reaction [3]. Certain constraints were included to ensure that the backbone structure reflects that of the protein crystal structure: both terminal backbone atoms were geometrically fixed in the crystal structure configuration, both dihedral angles are constrained to values from the crystal structure and throughout the classical molecular dynamic trajectories, and the hydrogen bond between N^δ–H of His and the carbonyl O of Asn was constrained at a distance of 1.8 Å. Without these constraints, the subsystem would likely rearrange into a structure that does not represent the intein C-terminal structure but does minimize the gas phase energy. By contrasting the effects of mutations, electronic structure properties at critical points were studied, including those at the purely quantum mechanical transition state.

4.1.2.2 Energetic results

For the N-protonation mechanism calculated with the tripeptide system, the computational energy barrier for the His–Asn–Cys system in the gas phase was 27.95 kcal/mol, in good agreement with the experimental results of ∼21 kcal/mol [24]. For a system roughly 30 atoms smaller, the previous gas phase energy barrier was ∼33 kcal/mol [28]. This difference indicates that even the most basic approximation of the tertiary structure is important for accurate prediction of certain reaction energy barriers, as we will see with the QM/MM reaction. Additionally, we have tested and confirmed that the hydrogen bond between N^δ–H of His and the carbonyl O of Asn (dashed line in Figure 4.2) caused O to not accept a proton from H_3O^+. This hydrogen bond is usually found at the C-terminus of inteins and is important for reducing the possibility of proton transfer to the carbonyl O. In fact, the normally highly exothermic reaction for H_3O^+ to donate a proton to the carbonyl O atom is endothermic for cases where O is hydrogen bonded with another group [80].

Table 4.1 summarizes the calculated energy barriers and relative rate constants for the gas phase tripeptide system with several His–Asn–Xxx mutations. By including additional atoms, the gas phase energy barrier with Xxx = Cys (27.95 kcal/mol) was less than the previously calculated barrier for a smaller system (33 kcal/mol [28]) due to polarity and geometrical effects. The larger system used here was expected to more closely match the experiment of 21 kcal/mol, which is does, because of the additional mechanical and electronic influences of nearby protein and solvent groups.

---

[3]Gas phase energy barriers are typically higher than barriers that include electrostatic contributions such as implicit solvent calculations.

Table 4.1: **Tripeptide energy barriers (ΔE) for various C-extein muta-**
tions (His–Asn–Xxx), percent change (%ΔE) from His–Asn–
Cys energy barrier, and expected change in reaction rate k_{rel}
compared to His–Asn–Cys. Structures were geometrically op-
timized with the B3LYP/6-311++G(d,p) level of theory. The
percent change in the energy barrier, %$\Delta E \equiv \frac{\Delta E_{Xxx} - \Delta E_{Cys}}{\Delta E_{Xxx}}$ *
100%. Reaction rates k are relative to the His–Asn–Cys wild-
type at $T = 310.15$ K (37 °C. The Arrhenius equation was used
to compare the relative reaction rates between two mutants:
$k = k_1/k_2 = e^{-(\Delta E_1 - \Delta E_2)/RT}$, where k_i and ΔE_i were the reaction
rate and energy barrier for the ith mutant, respectively; R was
the gas constant and T was the temperature in Kelvin.

Mutant (Xxx)	ΔE [kcal/mol]	%ΔE	k_{rel}
Cys	**27.95**	0.00	**1**
Thr	27.56	-1.39	1.88
Ser	27.75	-0.71	1.38
Ala	28.64	2.46	0.32
Val	28.97	3.64	0.19
Met	**29.58**	**5.83**	**0.07**

The energy barrier of the His–Asn–Met system was 1.63 kcal/mol higher than the His–Asn–Cys system, which corresponds to a 5.83% increase in the energy barrier. When Cys was mutated to Met, the relative C-terminal reaction rate was predicted to be 0.07 as fast, or decreased by more than an order of magnitude (14.0), which is consistent with experimental results [24, 26]. Interestingly, this model predicts that Thr and Ser instead of Cys will be slightly more effective at pH-dependent C-terminal cleavage, a prediction that is consistent with the +1 position being occupied by Cys, Thr, or Ser in nature, and will be tested in experiment. In the context of splicing, experiments have shown that Cys, Ser, and Thr are the only amino acids with the ability to complete the transesterification step of splicing [5], which is consistent because they also are the most efficient at C-terminal cleavage according to the calculations presented here.

4.1.2.3 Charge analysis

Natural Populations Analysis (NPA) [81] was used to study the electron population and the partial atomic charges. Figure 4.3A illustrates the effect of amino acid mutation on the scissile peptide bond distance and Figure 4.3B shows the sum of the NPA charges for
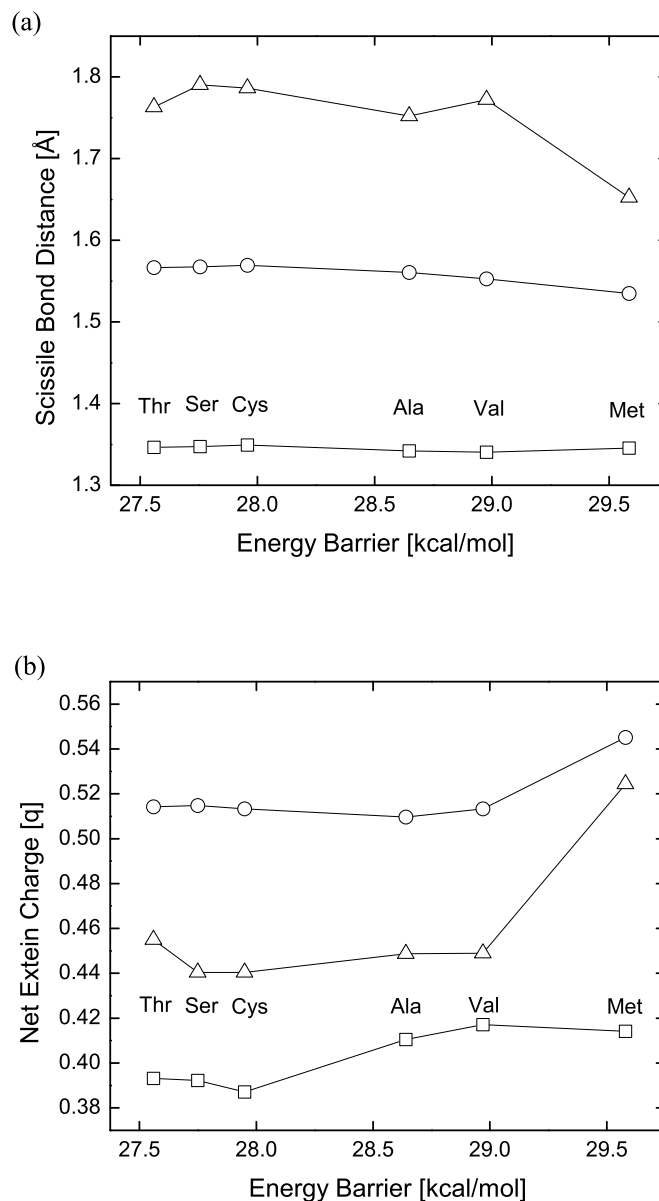
Figure 4.3: Relaxed scissile peptide bond distance (a) and NPA charges summed for atoms on the C-extein (b) for the tripeptide gas phase system, His–Asn–Xxx (Xxx = Thr, Ser, Cys, Ala, Val, Met). Both the scissile bond distance and the net charge for the C-extein amino acid (Xxx) are plotted as a function of the specific mutant's energy barrier and are shown for the normal amide, ($\square$); the N-protonated amide, ($\bigcirc$); and the Asn cyclization transition state ($\triangle$).

the mutated C-extein residue, starting with the -NH at the scissile junction and including the side chain. The scissile bond distance and charge results are shown as a function of each mutant's energy barrier, and include the normal amide, the N-protonated amide, and the transition state corresponding to the pH dependent C-terminal cleavage reaction. For the neutral amide, the C–N scissile peptide bond distance was 1.3492 Å for Cys, which decreased to 1.3455 Å for Met. Although this change was extremely small, it does confirm that the amino acid side chain played a small but perceptible role in the properties of a normal peptide bond (which is well known from proton exchange experiments [82]). For the N-protonation step and then the Asn cyclization transition state, the correlation between short scissile bond distance and high energy barrier was more apparent: a shorter peptide bond implied more π-bond resonance between C and N, less π-bond resonance between C and O, and more energy was required to break the C–N bond. An elongated peptide bond implied less π bonding between C and N and less energy necessary for peptide bond cleavage [72].

A correlation between the energy barrier and the net charge can be seen (Figure 4.3B), especially for the Cys/Met mutation, signifying that the residues that were able to accept more electrons exhibit a reduced energy barrier whereas the residues that were less likely or unable to accept electrons displayed an increased energy barrier.

4.1.3   Single amino acid molecules

4.1.3.1   Electron affinity and ionization potential analysis

To further elucidate the effect of the mutation of the first C-extein amino acid side chain on the energy barrier, the isolated Cys and Met amino acids were studied. The electron affinities (EA) and ionization potentials (IP) for each were calculated with the B3LYP/6-311++G(d,p) level of theory. The EA for Cys, (the amount of energy gained or lost when the system goes from neutral to negatively charged), was 6.79 kcal/mol. For Met, the EA was 8.27 kcal/mol, signifying that the side chain of the gas phase Cys residue was more electronegative than for Met. The reason that Cys was more stable with charge than Met was due to the bonding for each S atom. Although each side chain contained an S atom, for Cys the S atom was bonded to one methyl group and one H atom. For Met, both bonds of the S atom were to methyl groups, hence different electron occupation properties. In changing from neutral to negatively charged, the partial charge of S for Cys changed from -0.01051 to -0.11874 units of charge, corresponding to the

addition of 0.10823 electrons. For Met, the charge went from 0.16894 to 0.12532 units of charge, corresponding to the gain of only 0.04362 electrons. The S of Cys was able to accommodate more than twice the amount of delocalized electron population as compared to Met, indicating more energetic stability in the negatively charged system. The difference in ionization potential (IP) for the same isolated Cys and Met amino acids was calculated. The removal of one electron from Cys required 203.05 kcal/mol while that for Met was 191.14 kcal/mol. Combining the fact that Met was more stable when an electron was removed, and the fact that Cys was more stable when an electron was added, we conclude that the "electron pulling" and "electron pushing" properties of the first C-extein amino acid side chain must have an effect on the actual properties of the scissile peptide bond.

4.1.3.2 Energetic analysis of molecular orbitals near the Fermi energy

For the isolated amino acids (Thr, Ser, Cys, Ala, Val, and Met), the highest occupied molecular orbital (HOMO) for the neutrally charged system as well as the negatively charged system was compared. The difference in energy between the HOMO of the electron doped (negatively charged) and the neutral system is termed the energy gap, and is shown in Figure 4.4. From this analysis of the negatively charged amino acids (geometrically optimized with neutral charge), the isolated amino acids are ranked in order of the energy barrier found when they are the mutant for the tripeptide system, and there was a clear trend in the energy gap between the neutral and negatively charged molecules. The energy gap was closely related to the electron affinity of the molecule: as the energy barrier increased for a particular mutant, the gap decreased. This single amino acid analysis is of particular interest because from the electronic structure properties of an isolated molecule representing an amino acid side chain, calculated properties such as the electron affinity, the ionization potential, and the molecular orbital energy levels may explain and perhaps predict the relative reaction rate for an unknown mutant at the first C-extein position.

The localization of the EA densities found for molecules characterized in Figure 4.4 is plotted as a volumetric surface in Figure 4.5, which shows the difference in electron density between the neutral (optimized geometry) and negatively charged (single point geometry) single amino acid residues (Thr, Ser, Cys, Ala, Val, and Met). The presence of electrons on the molecular side chain was observed for amino acids that are more efficient when downstream of the scissile peptide bond in intein C-terminal cleavage.
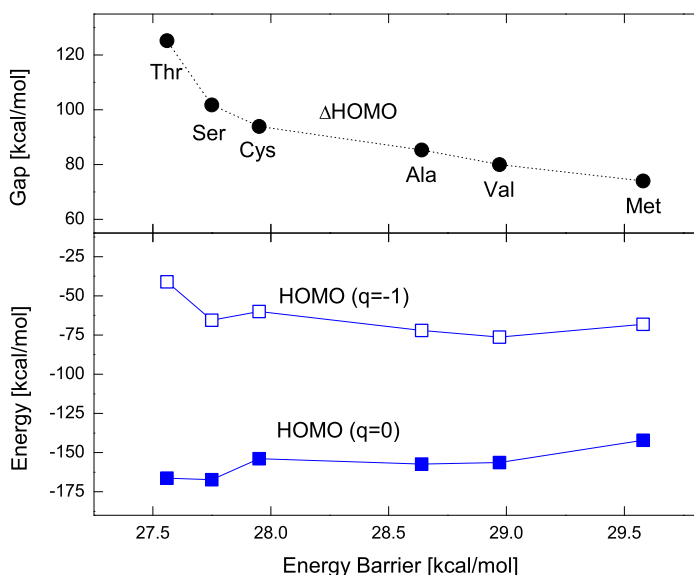
Figure 4.4: Energies for the highest occupied molecular orbital (HOMO) for the neutral system (■) and the negatively charged system (□) for the isolated amino acid molecules (Thr, Ser, Cys, Ala, Val, Met), shown in order of their energy barrier found independently for the tripeptide reaction calculation. The difference between these energies is the energy gap (●) and is clearly dependent to the energy barrier for the given mutant.

4.1.3.3 Tripeptide analysis

Returning to the tripeptide system shown in Figure 4.2, Table 4.2 shows electron population analysis for orbitals with $l = 1$ angular momentum ($2s$ orbital), as well as total occupation for $l = 0, 1$ ($2s$ and $2p$ orbitals). From the analysis of target atoms belonging to the scissile peptide bond, the expected differences in electron population between Cys/Met mutants were observed. Specifically, the N atom for Met was generally more occupied with electrons than Cys, which gave it a greater negative charge.

For both mutants, the N atom showed a considerable increase of $2s$ electrons, which corresponded to C and other atoms returning σ electrons to N when the C-N bond was elongated after N-protonation. A similar situation with σ electron back-transfer to N was found for peptide bond rotation, where at the transition state of 90° the N atom lost π electrons although there was an increase in σ electrons to N [72]; this phenomenon explains why N actually became more negative as similarly seen in the present study. The
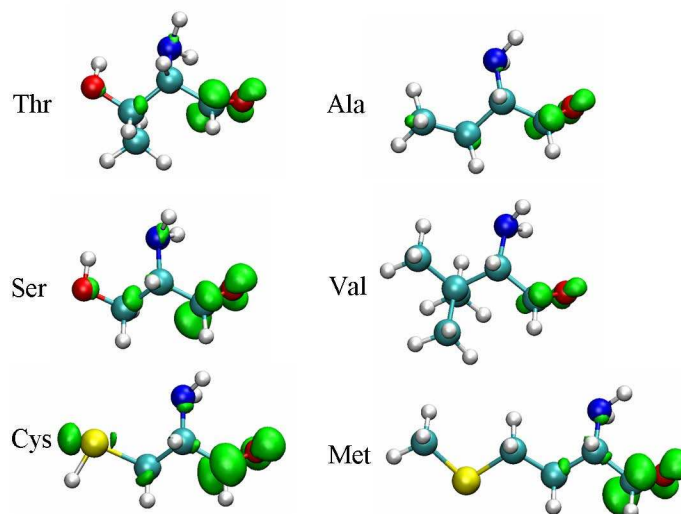
Figure 4.5: The electron affinity (EA) density for single amino acid molecules (Thr, Ser, Cys, Ala, Val, and Met). The electron density surface describes the delocalization of the electron affinity when an electron is added to the system, thus going from neutral to negatively charged ($\Delta\rho$). For downstream amino acids that were efficient at C-terminal cleavage (Thr, Ser, Cys), the EA density extended to the side chain. For amino acids that were less efficient (Ala, Val, Met), the EA density remained on the peptide-like part of the molecule, and away from the side chain. Atom colors are as follows: carbon is cyan, nitrogen is blue, oxygen is white, sulfur is yellow, and hydrogen is white; the electron density surface is green [83].

$2p$ orbitals for N showed distinct differences for the Cys/Met mutations – even for the neutral ground state which was a normal amide system, a distinction that signified the side chains of adjacent amino acids were important in dictating the exact properties of the peptide bond.

For the normal amide, the charge of the peptide N for Cys was -0.616 and for Met the charge was -0.641. For the N-protonation case, the charge of N for the Cys case was -0.660, where for Met the charge was -0.710. For the transition state, the charge on N for Cys was -0.684, and for Met was -0.699. For all three cases the charge of N for Met was more negative than for Cys, which was consistent with the electron affinity calculation described previously. The side chain plays a subtle yet important role in the electrostatic environment during the cleavage reaction. By having less charge on N, the -NH$_2$ group is

Table 4.2: **Atomic orbital populations for the** $2s$ **and net** $2p$ **orbitals as well as the total electronic occupation for the peptide N atom in the gas phase tripeptide calculation. N is generally less occupied by electrons for Cys as compared to Met, which is consistent with single amino acid electron affinity results. The sum of electron occupation for the** $2p_x$, $2p_y$, **and** $2p_z$ **orbitals is written as** $2p$**. The NPA charge is calculated by subtracting the total electron occupation from the atomic number; a larger electron occupation signifies a more negative charge.**

Orbital	Mutant	Occupation		
		Neutral ground state	N-Protonated	Transition state
[$2s$]	Cys	1.250	1.359	1.386
	Met	1.259	1.360	1.376
[$2p$]	Cys	4.341	4.285	4.277
	Met	4.357	4.329	4.299
Total	Cys	7.616	7.660	7.684
	Met	7.641	7.710	7.699

more energetically favored to leave. From this electron population analysis, differences in the electronic structure of the scissile peptide bond for Cys and Met were observed, which explained why the energy barrier for Cys and Met mutants would be distinct despite an identical mechanism.

4.2 Reaction analysis with QM/MM calculations

The full protein QM/MM reaction profile was initially calculated with the QM active site region of His–Asn–Cys, and two water molecules (2346 protein atoms, 4161 water atoms, and total 53 QM atoms) [84]. Figure 4.6 shows the QM/MM energy barrier with and without electrostatic embedding. The energy barrier was 24.96 kcal/mol for the QM/MM calculation with geometry optimization, in excellent agreement with the 21 kcal/mol measured experimentally [24].

4.2.1 Effect of mutation on energy barriers

The energy barrier difference for the Cys/Met mutation is of interest in the context of a QM/MM calculation, but because the Met side chain was too spatially extended to simply replace the smaller Cys side chain, additional classical MD simulations were

(a)

(b)
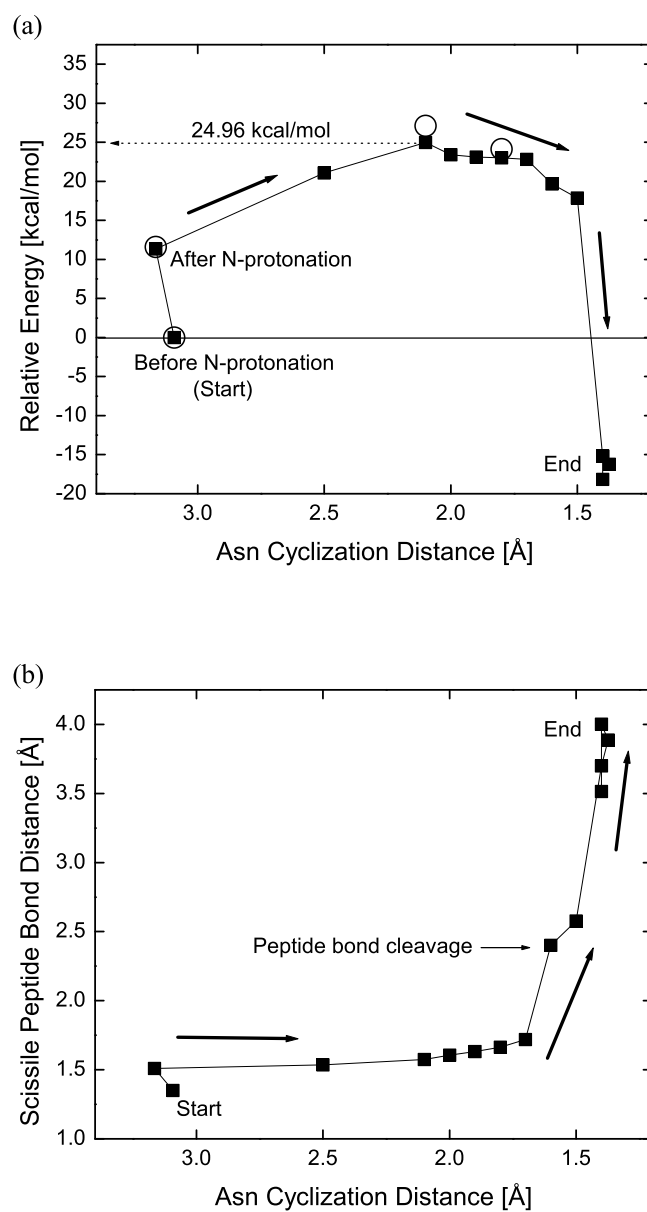
Figure 4.6: Combined QM/MM reaction energy profile (a) and distance of the scissile peptide bond during breakage (b) for His–Asn–Cys plus two water QM system. QM/MM geometry optimization, (■). QM/MM + charge embedding single point energies, (○).

performed (starting from the initial intein plus extein structure) but with Met at the C-extein +1 residue. Once the full protein system was equilibrated, the QM active site was partitioned to be His–Asn–Met plus the two water molecules in the same location as before (59 total QM atoms). The Asn cyclization reaction coordinate was scanned after N-protonation by H_3O^+. To compare the effect of the Met/Cys mutation directly, the smaller Cys was substituted for Met, and the geometry was again relaxed. By doing this, the change in reaction energies may be compared directly because the original protein structures were common for both Met and Cys residues.

These structures were in near total overlap, with the exception of the side chain of the (+1) amino acid, either -CH$_2$-SH for Cys, or -(CH$_2$)$_2$-S-CH$_3$ for Met. Using the B3LYP/6-31G(d,p) level of theory, independent reaction profiles for the Met/Cys mutation were calculated. For Met the barrier was 27.07 kcal/mol and for Cys was 26.17 kcal/mol. The His–Asn–Met QM active site (as part of the QM/MM system) had an energy barrier of 0.90 kcal/mol higher than His–Asn–Cys, which corresponded to ratio between reaction rates of $k = k_{Cys}/k_{Met} = 0.22$, in good agreement with experimental results and consistent with the tripeptide system conclusions [24, 26].

4.2.2 Effect of mutation on electron occupation

In addition to energy barriers, the Mulliken charge [85] was calculated for critical atoms[4]. For the N atom of the scissile bond and for the ground state, the partial charge was -0.538 for Cys and for Met was -0.545. For the N-protonation state the partial charge of N was -0.609 for Cys and was -0.615 for Met. At the transition state, the charge for Cys was -0.584 and for Met was -0.598. In all cases the partial charge of the N atom for the Met mutant was more negative, which was consistent with the tripeptide results, and is explained by using the electron affinity and ionization potential for the isolated Cys and Met amino acids. When the net Mulliken charge was summed for the C-extein residue (Cys or Met) in the QM/MM context for the normal amide ground state, for Met the net charge was 0.225, and for Cys the net charge was 0.209.

Within the QM/MM system, the charge for the backbone and side chain of the first C-extein residue was added. The net charge of Cys was more negative than Met, which is in agreement with the model QM calculations described in the preceding paragraphs.

By combining model system QM calculations and full-protein QM/MM simulations,

---

[4]Natural Population Analysis (NPA) is not implemented with QM/MM at this time.

the non-mechanistic regulation of reaction rate regulation for single amino acid mutations near to the active site was confirmed, explained, and predicted. Similar methods are also useful for testing an unknown mechanism based on the correlated experimental results of kinetic data (from non-essential amino acid site-directed mutagenesis).

4.3   Conclusions: C-terminal cleavage

The C-terminal cleavage reaction and the previously proposed N-protonation mechanism were tested by increasing the QM system size by 30 atoms to at least 53 atoms. In addition, full-protein QM/MM analysis was performed. The pH dependent C-terminal cleavage reaction undergoes simple proton-catalysis by a hydronium ion that protonates the peptide N atom. The peptide bond, now resonance destabilized, is elongated and the peptide C atom is open for attack by the Asn side chain. During Asn cyclization, the peptide bond cleaves while an aminosuccinimide ring is formed. The final step involves the donation of the extra proton on the aminosuccinimide to the $-NH_2$ leaving group *via* water, thus making the leaving group positively charged. Our QM/MM results included the effects from the protein interior, both mechanical and electrostatic.

The "non-mechanistic" role of the first amino acid of the C-extein was confirmed. This amino acid, although not necessary for C-terminal cleavage, did have an effect on the reaction rate by about an order of magnitude, as measured by Wood *et al.* [19, 24, 26]. In this study, the precise energy barrier for C-terminal cleavage (and hence reaction rate) was shown to be dependent on the side chain of the amino acid downstream from the scissile bond. Explained by the electron occupation and partial atomic charges for each residue at the C+1 position, considerable differences that led to a distinction in energy barriers were calculated and found to be in agreement with experimentally observed reaction rates.

CHAPTER 5
SPLICING

5.1   Splicing introduction

During the process of protein splicing, an intein auto-catalytically cleaves both the N-terminus and C-terminus and simultaneously ligates the flanking peptides (exteins) [5]. Inteins are protein segments that catalyze splicing in a host protein; specifically, the N- and C-exteins that flank the intein[1, 2]. From random and site-directed mutagenesis, the splicing reaction has been proposed to depend on several highly conserved residues [19, 20, 21, 22] located in non-gapped conserved regions (blocks) that are consistent for most canonical inteins. Conserved residues are highlighted in Figure 5.1. Despite this information, an atomic-level description of the splicing mechanism is not available.

Intein crystal structures [11, 12, 13, 14, 15, 16, 17] have shed light on the splicing and C-terminal cleavage mechanisms. One limitation in using crystal structures for mechanistic studies is that they are mostly product proteins that do not include exteins, which were excised during the splicing reaction. Additionally, there is often increased spacing between N- and C-termini of the intein due to electrostatic interactions with solvent or metal ion binding. In general, the intein termini of crystal structures are considered to be geometrically flexible. Despite these concerns with intein crystal structures, they are critical for understanding many aspects of intein reactions. One intein (*Ssp* dnaE) was crystallized both as a product protein as well as an engineered inactive precursor, with exteins [17]. In this case, short exteins were crystallized with the intein, and the catalytic residues had to be substituted in order to inhibit the reaction. Therefore, the exact state in which the inteins are found before splicing is not clear since to observe the precursor, mutations that inhibit splicing need to be engineered. Nevertheless, overlap between inhibited precursor and spliced product was reported for the protein backbone.

Experiments on the splicing reaction are exceedingly complex due to the various possible products. For example, to measure the first step of splicing, namely the N-terminal thioester formation, a typical construct is used where the C-extein is not present (therefore no C-terminal cleavage is possible). Dithiothreitol (DTT) is added in order to cleave the N-terminal thioester, although the exact process is not precise because without the C-extein, there is a proximal charged and flexible C-terminus which may affect the
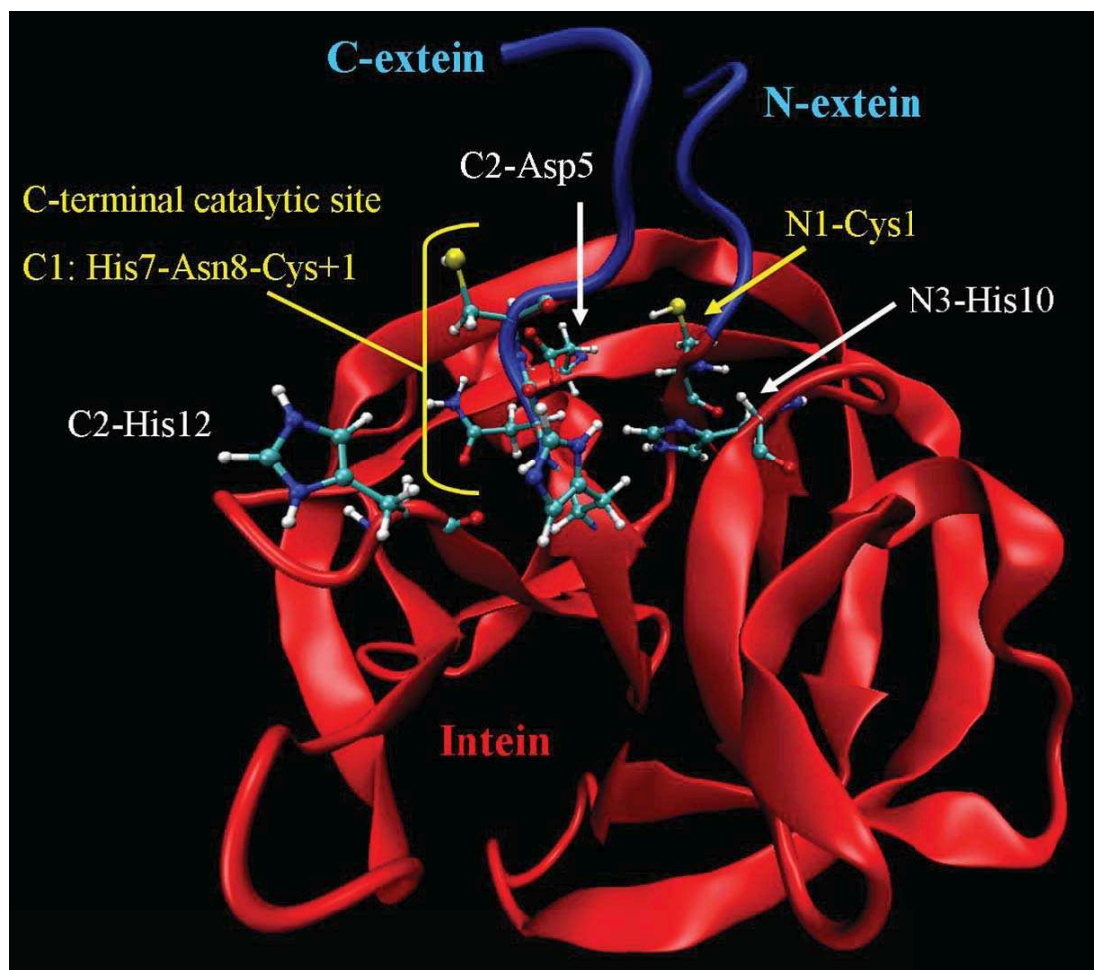
Figure 5.1: The conserved amino acids treated as the quantum mechanical active site are highlighted in ribbon diagram of the the crystal structure.

reaction steps and rates at the N-terminal. Because of this, experiments using DTT and measuring N-terminal cleavage most likely encounter different chemistry from N-terminal cleavage within the context of splicing.

5.1.1 Splicing reaction

The intein substrate is naturally tethered to its active site. This makes the catalytic mechanism different than typical enzymes because there is binding between substrate and enzyme, and here the turnover number equals one. Because the substrate is tethered to the active site, it has been extremely difficult to isolate molecular inhibitors for the splicing reaction [86]. From *in vitro* experiments, zinc and copper atoms are known to inhibit

splicing [87, 88]. Due to the inability of zinc to enter the *E. coli* cell, *in vivo* experiments with precursor or intermediary intein constructs inhibited by zinc are not likely to occur.

Because of these factors, the splicing mechanism is not understood on the atomic level [89], and this has limited the biotechnological applications utilizing inteins [90]. Splicing is based on a four steps, and it is hoped that elucidation of the intein splicing mechanism will lead to new schemes to either trigger the reaction or to inhibit it. Biotechnological applications include switches and delivery devices for bioseparations [6, 7], drug development [8], and molecular sensors [9, 10].

To date, the splicing reaction scheme, based on crystal structures [11, 12, 13, 14, 15, 16, 17] and mutagenesis information, is as follows: The thiol group (-SH) of the first intein residue (N1-Cys1) is involved in an N-S shift, where it attacks the downstream carbonyl carbon thus releasing the peptide nitrogen. This intermediate state is called the N-terminal thioester. After N-terminal thioester formation, transesterification occurs. In this step, the C-terminal Cys (first residue of C-extein and not present in the crystal structures) attacks the carbonyl of the thioester, thus combining the N- and C-exteins. Once transesterification is complete, the C-terminal Asn residue undergoes cyclization into a succinimide and C-terminal cleavage occurs. Now, the N- and C-exteins are fused and the intein is released. The ligated product (N- and C-extein) then undergoes the final reaction which involves an S-N shift, replacing the thioester with a more stable peptide bond. Interestingly, the two Cys residues may be replaced by Ser[5] with similar reactions steps, although the reaction usually occurs at a slower rate due to decreased acidity of the Ser OH side chain compared to the SH of Cys [4].

The reaction scheme described above does not include the explicit role and pathway of protons. Prior to N-terminal thioester formation, the thiol group must lose its proton and become a thiolate (-S$^-$). Also, prior to transesterification, the C-extein thiol must similarly be ionized. To be an adequate leaving group, the peptide nitrogen must have two protons (-NH$_2$). To be a stable N-terminus, at biological pH levels the nitrogen will have three protons (-NH$_3^+$).

5.1.2 Effects of mutation at highly conserved locations

The limited understanding of the role of protons is based from mutagenesis experiments [19, 20, 21, 22] and X-ray crystallography studies [11, 12, 13, 14, 15, 16, 17]. The

---

[5]The C-extein Cys residue can also be replaced with Thr

splicing reaction has been proposed to proceed *via* the following highly conserved residues for canonical inteins, which are shown in Figure 5.1:

- N1-block Cys (N1-Cys1), which undergoes an N-terminal N-S acyl shift. This results in the N-terminal thioester. Ser or Thr may be used in certain constructs with a noticeable impact on reaction rate.

- N3-block His (N3-His10) regulates N-terminal hydrolysis and is the most conserved of all intein residues.

- C2-block Asp (C2-Asp5) links the N- and C-termini at the active site. This coupling leads to successful splicing and not independent N- and C-terminal cleavage events. When mutated to Gly, the intein is converted to the C-terminal cleavage mutant (CM), a highly efficient and pH dependent mutant [19, 24, 28, 29]. One major question exists: does CM have inhibited splicing or is cleavage just more efficient? A catalytic role for C2-Asp5 would explain why this residue is necessary chemically.

- C2-block His (C2-His12) may be a proton acceptor/donor for C-terminal cleavage reaction in the context of splicing.

- C1-block C-terminal catalytic site and first C-extein residue (C1: His7 - Asn8 - Cys(+1)) are important for C-terminal cleavage. The penultimate His side chain has a highly conserved hydrogen bond with the carbonyl oxygen, which is considered to be structurally important. One intein lacks the penultimate His and uses an Arg residue at a different position for a similar role [17]. The role of Asn and the C-extein Cys are described in the C-terminal cleavage chapters.

- An important yet not understood mutation is N3-Val4Leu (V67L), which converts the deletion variant mini-intein into a faster splicing mutant (SM).

Extein sequence, too, may play a role in intein reactions [12, 14, 17]. The importance of conserved residues in the exteins was shown for split inteins, specifically for residues near the scissile peptide bond at both the N- and C-termini [91]. Split inteins undergo trans-splicing, where the intein is split into an N-intein and C-intein that are connected to the N- and C-exteins, respectively. In this case, the separate intein fragments must come into contact and enter a conformation that is conducive to splicing, and the exteins may play a role in this conformational search.

Here we present the intein splicing mechanism with an atomic-level description based on first principles density functional calculations.

5.2   Description of model system

The splicing mutant (SM) crystal structure (PDB code 2IMZ) was computationally amended to include both an N- and C-extein. The solvated system was equilibrated with classical molecular dynamics (MD) simulations with exteins present. To test the N-S shift, QM/MM calculations were performed. The protein consisted of 2351 atoms and there were 1321 water molecules present, for a system total of 6314 atoms. The N-terminal active site is based on the N-extein residue Lys(-1) and the intein residues N1: Cys1-Leu2. Also included are N3-Thr7, N3-His10, C2-Asp5. From the C1 block and C-extein (C-terminal active site): Val6-His7-Asn8-Cys(+1)-Ser(+2), where the peptide bond between C1-Asn8 and the C-extein Cys is cut during C-terminal cleavage. For non-mechanistic residues, those that affect the QM system *via* polarization, the entire amino acid may not be chosen for inclusion in the QM system – an example of such a residue is N3-Thr7, where only the side chain is included. Nine water molecules were included explicitly in the quantum mechanical calculations. Because of the extremely large QM active site (130+ atoms), it should be noted that for various calculations this QM system is either appended or truncated[6], but for all energy calculations and comparisons, the systems are directly identical in all atomic constituents (number of electrons, charge, spin, and nuclei).

5.3   First principles splicing mechanism

Protein crystal structures allow for mechanistic predictions based on structural information. Random and site-directed mutagenesis studies are helpful for mechanistic prediction because they show the reaction may be inhibited, slowed, or sped up through mutation. Inteins are a unique protein because they can be thought of as their own enzyme, where the exteins can be thought of as the substrate. The mechanistic details for their auto-catalytic behavior are coded in their amino acid sequence. Because reactive intein crystal structures do not include exteins (with the exception of inteins engineered to be inhibited) and because exteins are an important part to the mechanism, structural

---

[6]The absolute energy is dependent on many factors including densinty functional, basis set, system size, and polarization environment. Because the splicing reaction occurs over a large portion of protein, it is necessary to modify the active site model system for different steps along the reaction path. All energies presented are relative to analogue systems.
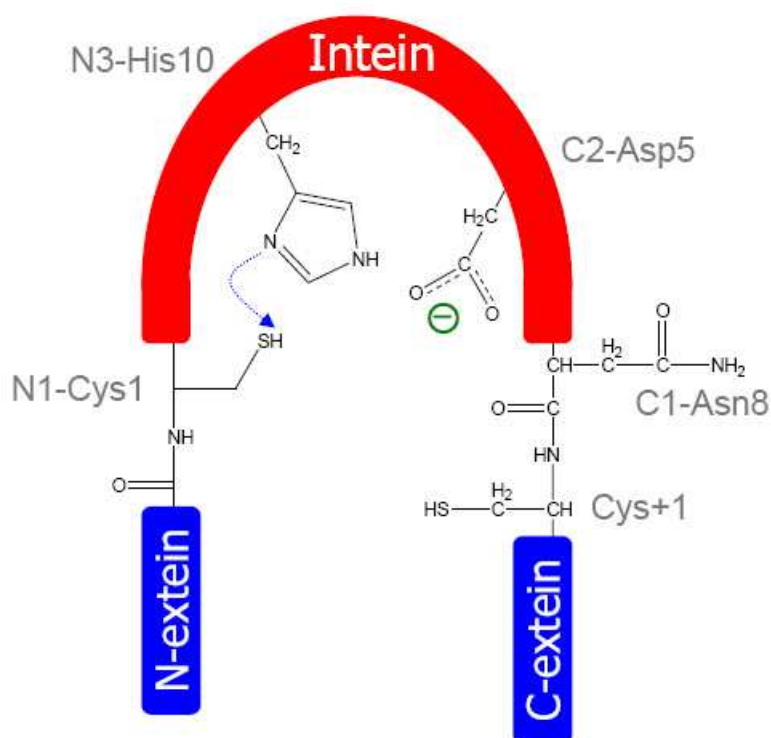
Figure 5.2: **Ground state for the intein splicing mechanism. Arrows indicate the direction of electrons. Critical residues are shown, water molecules included in the calculation are not shown for clarity.**

data are not able to provide complete mechanistic understanding of the splicing reaction at the atomic level. Indeed, in terms of protons, which are the driving force of chemical reactions and acid/base chemistry, splicing is not understood. Instead, certain amino acids are determined to be important or critical for splicing. We present the splicing mechanism that is based on first principles calculations and that utilizes those amino acids that are highly conserved.

5.3.1 N-terminal thioester

The ground state of the intein splicing system is shown in Figure 5.2. The proposed first step of splicing is the N-S shift, where the N1-Cys1 thiol group (-SH) is ionized and this thiolate group then attacks the carbonyl C of the upstream peptide bond. The newly formed thioester C(=O)-S bond causes the electrons shared between C and N (the peptide bond) to be unshared and the C-N bond breaks. For this step, the peptide N atom requires
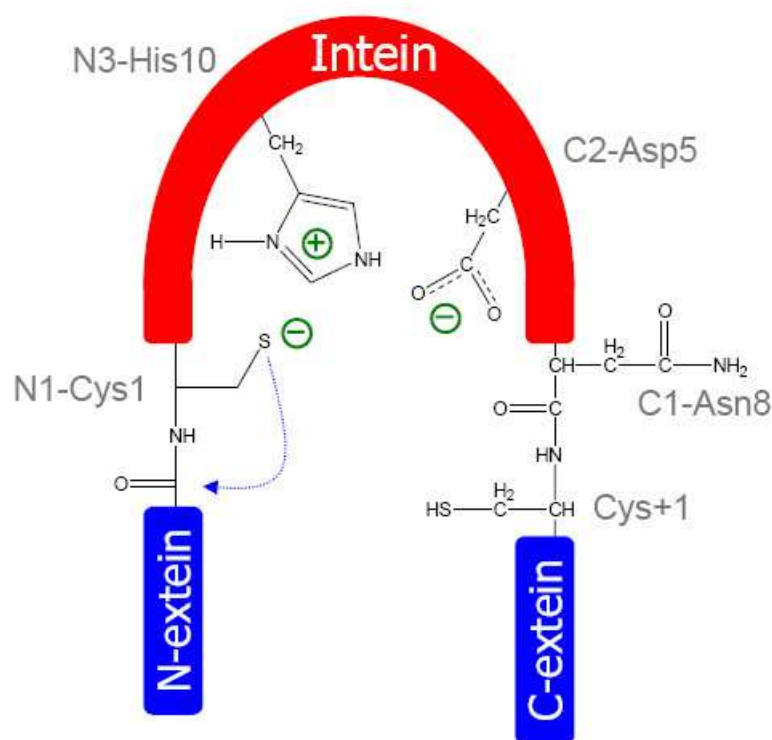
Figure 5.3: The N1-Cys1 side chain is ionized to a thiolate and gave its proton to N3-His10.

protonation (-NH$_2$ terminal) in order to be a sufficient leaving group.

The role of N3/His10 is first as a base and then as an acid. This residue, when in its neutral configuration, is in position to accept a proton from the N1-Cys1 thiol group, *via* water. This step in the splicing reaction is shown in Figure 5.3. The energy for this proton transfer (N1-Cys1 to N3-His10) is an upper limit for the N1-Cys1 ionization energy and was computationally found to be 19.27 kcal/mol. The ionization energy is a computational maximum because the thiol group is not likely to exist as a charged group without being somewhere in the process of the N-S shift reaction, and the protein was not equilibrated to accommodate this proton transfer. Once the proton has moved from N1-Cys1 to N3-His10, and during the thioester formation, the N3-His10 residue, now positively charged, is in position to donate a proton to the scissile N-terminal peptide nitrogen (without the need for a transitory water molecule).

The energy difference between the system where N3-His10 is positively charged and the system where there is a terminal NH$_2$ on the newly formed thioester is 25.05 kcal/mol. These energy values are approximate due to the static nature of these calculations, which

does not allow for full-protein equilibration of the structure with the intermediate N1-Cys1 thiolate and positively charged N3-His10. There exists an additional intermediate state that has an energy of 18.95 kcal/mol, where a ring structure is present and N is singly protonated and the carbonyl O is also protonated. Although this state does not show bond breakage between C and N, it does likely undergo C(=O)-S cleavage with DTT despite not necessarily being on the splicing reaction path. If this state were experimentally realized and more stable than the intermediate structure on the reaction pathway, then N-terminal cleavage *via* DTT may occur independently and outside the context of splicing.

From these calculations, the N-terminal thioester formed due to the combined acid/base properties at near-neutral pH levels of the N3-His10 residue, which first accepts a proton from solvent and then donates a proton back to the N-terminus.

NH_2 termination is necessary for an N-terminal leaving group, but is not optimally stable. Based on typical pK_a values, at neutral pH, solvent exposed NH_2 terminal groups will be again protonated and positively charged ($-NH_3^+$). The average pK_a for a solvent exposed N-terminus depends on the downstream side chain and ranges from 8.8 to 10.8 [73]. For this reason, an additional proton was included in the system (charge and spin are constant input parameters), and was tested at several positions (N3-His10 being one of them). With the thioester present, a stable state was found for $-NH_3^+$ termination, which was important for understanding the splicing reaction at the atomic level.

The spacing between the termini was too distant for the C-extein Cys+1 residue to attack the N-terminus. This could be because the crystal structure is the product protein segment, and the ends are flexible, charged, solvent exposed, and possibly metal-chelated. For this reason, the N- and C-terminal active sites were adiabatically combined after N-terminal thioester formation using both fully classical and semi-empirical/classical simulations (QM/MM). After equilibration, a stable minimum was found with first principles methods; no constraint was necessary between the N- and C- termini with spacing of approximately 4.5 Å, which is sufficiently close for the transesterification step.

The intermediate structures described here are accessible only with simulations. With these, the atomic-level details of the splicing mechanism were observed. In particular with this intermediate state (N-terminal thioester), the N-terminal $-NH_3^+$ group was in direct contact with both the highly conserved C2-Asp5 and the Cys+1 thiol group, which turned out to be an essential step in the splicing reaction. We will call these three groups the catalytic triad (N-terminal $-NH_3^+$, C2-Asp5 $-COO^-$, C-extein Cys+1 $-SH$), shown in
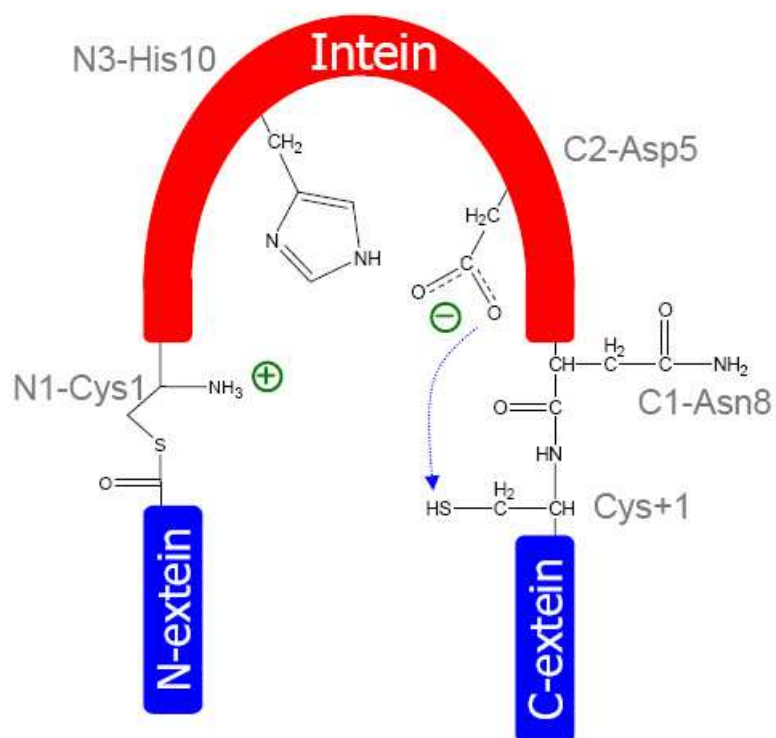
Figure 5.4: **N-terminal thioester shown after protonation of -NH$_2$ (-NH$_3^+$) from solvent, which is in electrostatic contact with Cys+1 and C2-Asp5 and catalyzes the transesterification step.**

Figure 5.4.

5.3.2 Mutation of C2-Asp5

The C2-Asp5Gly mutation[7] was experimentally concluded to inhibit splicing activity and enhance C-terminal cleavage activity, the latter having an improved reaction rate with low pH [19, 24]. The presence of Gly was proposed to have one of two roles. First, the smaller volume of Gly may allow solvent to access the C-terminal active site which would enhance C-terminal cleavage and show strong rate dependence on the pH of solvent. Second, the freedom of additional torsional rotation of Gly residues might allow the folded protein more flexibility for C-terminal cleavage. Neither of these hypotheses explained that splicing was inhibited with the C2-Asp5Gly mutation. Our experimental mutagenesis results indicate that when Ala or Asn is present instead of C2-Asp5, splicing is not observed

---

[7]The mutation formerly known as D422G.
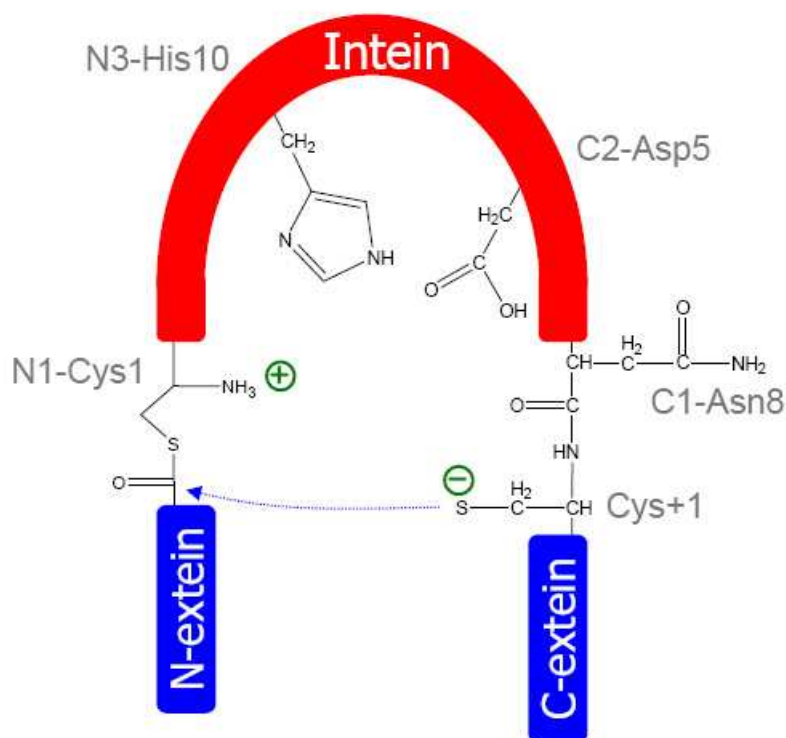
Figure 5.5: **The C-terminal Cys+1 experiences electrostatic repulsion due to the N-terminal -NH$_3^+$, and therefore donates a proton to C2-Asp5.**

and C-terminal cleavage is enhanced (although still less rapid than with Gly) [11, 23]. From the fact that splicing is observed with Asp or Glu, we can conclude that the side chain at position 5 in the C2-block is chemically/mechanistically involved in the splicing reaction and actually essential as a proton acceptor and donor.

Using first principles methods, the role of the C2-Asp5 was determined. The Cys+1 thiol group (SH) was in close contact with both the C2-Asp5 and the N-terminal -NH$_3^+$ group. It was observed that the Cys+1 SH group was in direct contact with the N-terminal NH$_3^+$ group, which is an unfavorable electrostatic interaction between (partially) positively charged hydrogen atoms (see Figure 5.5). From this, the thiol (-SH) group donated its proton to the carboxylate group of Asp. The triad was stabilized by having -S$^-$, -NH$_3^+$, and -COOH, in direct contact.

The C-terminal Cys+1 thiol group was located in a position in contact with the -NH$_3^+$ group that was newly formed after N-terminal thioester formation. This electrostatic interaction was unfavorable, and Cys+1 donated its proton to C2-Asp5, which goes

from negatively charged to neutral. The thiolate ($-S^-$) was then activated for the transesterification step.

5.3.3 Decreased stabilization of $-NH_3^+$ with C2/Asp5Gly

In addition to the catalytic role of C2-Asp5, we have studied the effect of Gly at that position. It was observed (see Figure 5.6) that with the C2-Asp5Gly mutation, the Cys+1 thiol group (-SH) is unable to donate a proton to the chemically inert Gly, as it did with Asp. Because of the need for electrostatic stabilization, it was observed that the $-NH_3^+$ terminal group spontaneously gives its proton to N3-His10, and becomes the non-charged $-NH_2$ group. This is a step backwards in the reaction, and confirms that splicing is impossible with the C2-Asp5Gly mutation. In addition, the presence of Glu (which is chemically similar to Asp) in that position results in some splicing products (albeit reduced reaction rate), consistent with this mechanistic assertion.

At this reaction step, C2-Asp5 is protonated, despite the typical pK_a of the aspartate side chain being 3.9 [73]. At biological pH levels, the side chain should be negatively charged (COO-). We have shown that due to the electrostatic effect of the $-NH_3^+$ terminal group, the Cys+1 thiol donated a proton to the Asp group thus making it a carboxylic acid (COOH), despite the biological pH level. This is important for the upcoming reaction step of C-terminal cleavage, which occurs after transesterification, and was shown independently to occur more rapidly at sub-biological pH levels [19, 24].

5.3.4 Transesterification

Transesterification occurs when the Cys+1 thiolate attacks the carbon N-terminal thioester (see Figure 5.7). This reaction step bridges the N- and C-exteins and breaks the bond between the N-extein and the intein. Most likely because the crystal structure of the product protein does not include exteins, there is a spacing of approximately 8 Å between N- and C-termini. Because intein splicing does occur, it can be concluded that the spacing between N- and C-termini will increase in the crystal structure compared to the precursor due to charged end groups, solvent exposure, and possibly metal binding [11, 12, 13, 14, 15, 16, 17]. After N-terminal thioester formation, the activated C-extein Cys+1 thiolate was capable of attacking the N-terminal carbonyl, thus completing the transesterification step and fusing the N- and C-exteins. The C-extein and intein remain chemically bonded until the C1-Asn8 cyclization reaction.
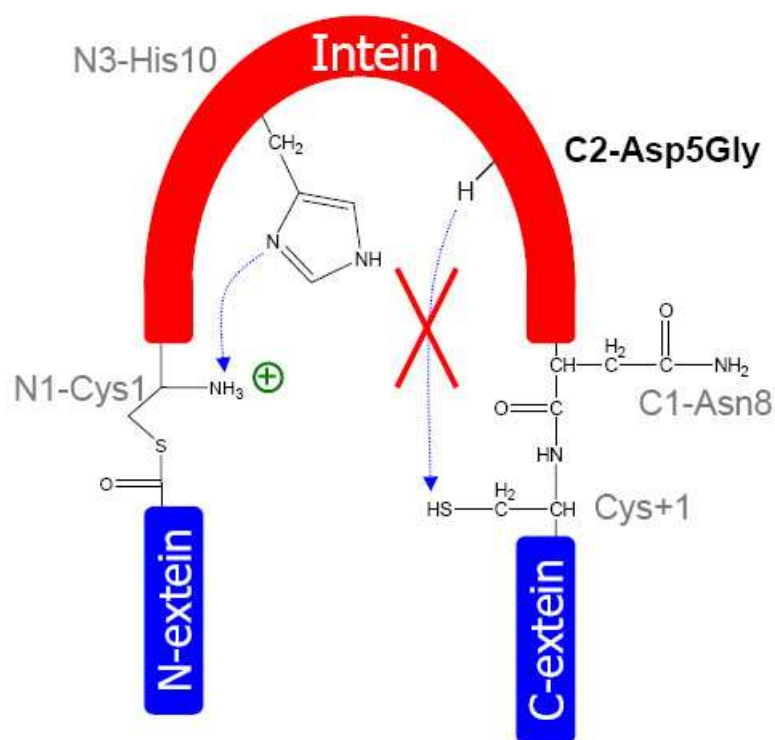
Figure 5.6: **With the C2-Asp5Gly mutation (D422G), the C-terminal Cys+1 cannot donate its proton to the C2-Asp5 side chain. To stabilize the unfavorable electrostatic interaction between Cys+1 the the N-terminal -NH$_3^+$, the N-terminal group gives a proton back to the N3-His10 side chain. This result indicates that the cleavage mutant (CM) is unable to complete the transesterification step of splicing and that the overall splicing reaction is inhibited. N-terminal hydrolysis should still be possible, although it may occur much less frequently than pH enhanced C-terminal cleavage.**

At this point, both the Cys residues (N-terminal N1-Cys1 and the C-extein Cys+1) have been independently ionized and have undergone a rearrangement reaction as a thiolate. Shingledecker *et al.* have shown that also for the *Mtu* recA intein [4], the pK$_a$ for the N1-Cys1 was 8.2, which is comparable to average values for the Cys side chain [73]. For Cys+1, the pK$_a$ was 5.8, which is considerably lower and would indicate the increased probability of finding Cys+1 as a negatively charged thiolate. The pK$_a$ of N1-Cys1 is not necessarily important for the splicing mechanism because the entire reaction will wait for this initial step to commence. However, the low pK$_a$ for Cys+1 is of extreme inter-
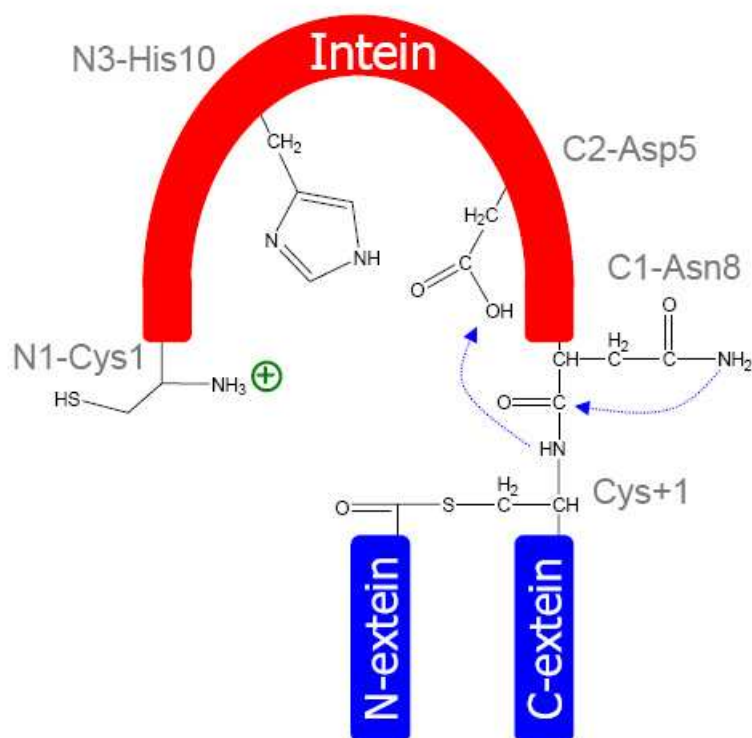
Figure 5.7: **Joining of the N- and C-termini *via* transesterification. This reaction step is catalyzed once the C-terminal Cys+1 is a thiolate, which depends on the pK_a of the side chain (discussed in text).**

est, because the increased probability of finding a thiolate at Cys+1 indicates that there is some electrostatic explanation why the side chain is so likely to be ionized, which is consistent with the mechanism proposed in this work. The Cys+1 attacking group must wait for the N-terminal thioester, which is much less stable than a peptide bond. The low pK_a indicates that the side chain ionization should be an energetically easy process, which should facilitate transesterification.

5.3.5 C-terminal cleavage

After transesterification, the subsequent reaction is C-terminal cleavage. In this reaction step, C1-Asn8 undergoes cyclization into a succinimide ring. Intein sequences and crystal structures indicate that either penultimate C1-His7 residue or a C2-block Arg residue is necessary for stabilizing the oxyanion hole during the C-terminal cleavage step within the splicing reaction [17, 92, 93, 94]. A hydrogen bond between the side chain -NH

of either C1-His7 or N3-block Arg and the C=O of the scissile peptide bond is highly conserved in crystal structures, although it is not expected to be proton transfer from this residue that catalyzes C-terminal cleavage.

Based on previous experimental mutagenesis studies [19], splicing may be inhibited and C-terminal cleavage may be isolated by a C2-Asp5Gly mutation, for which the role of Asp or Gly is explained in the preceding section. The C2-Asp5Gly mutant, termed the cleavage mutant (CM), also exhibited a higher activity in low pH, for which a mechanism was proposed in earlier chapters. Within the context of splicing, no striking pH dependence was observed, and although C-terminal cleavage still occurs, it occurs as a final reaction step and after the splicing reaction described above. It is an extraordinary feature that C-terminal cleavage, which is extremely rapid in the CM, delays itself until after the completion of splicing's N-terminal reaction steps. This can be explained by returning our focus to C2-Asp422, which was integral as a proton acceptor in triggering transesterification and the overall splicing reaction rate.

After transesterification, there is a proton on C2-Asp422 making the amino acid side chain neutral in charge (COOH). Average and solvent exposed pK_a values for aspartic acid are $\sim$3.9, indicating that at pH levels considered in splicing are much higher than typical average Asp pK_a values. The presence of the additional proton makes C2-Asp422 an acid (aspartic acid) and since it is protonated, an effect of a low pH environment is obtained, despite the overall biological pH level. The "effect" of a low pH environment is important for C-terminal cleavage and consistent with experimental results. The COOH group on the side chain, which was only formed after the N-terminal N-S shift, acts as a catalytic acid and donates a proton to trigger the C-terminal cleavage reaction, the first truly irreversible step in the splicing reaction mechanism.

5.3.6   S-N shift (finishing reactions with succinimide hydrolysis)

After C1-Asn8 cyclization triggered by proton donation from C2-Asp5, a succinimide ring is formed and the ligated N- and C-exteins are released. Two finishing reactions are expected (see Figure 5.8. First, the branched thioester undergoes an S-N shift and rearranges to form a peptide bond, which is energetically more stable. Second, succinimide may undergo hydrolysis and which would create a typical C-terminus (-COO$^-$) and a normal Asn or an iso-Asp residue, although in some crystal structures succinimide is observed. For example, the final position of the intein in the *Mtu* recA crystal structure
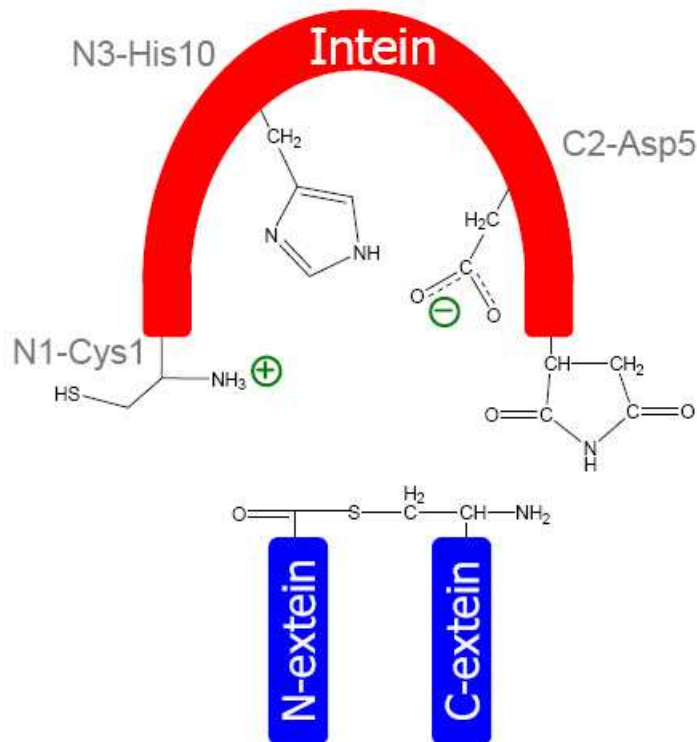
Figure 5.8: **After C-terminal cleavage and before the finishing reactions, the N- and C-exteins are ligated and the intein is released from the host protein. The finishing reaction will replace the thioester with a peptide bond.**

[11] was observed in the crystal structure to be either a succinimide [67, 95] or a negatively charged carboxyl group. The presence of succinimide gave insight to possible mechanisms for the isolated C-terminal cleavage reaction [28, 29].

5.4 Conclusions: Splicing

The mechanism for protein splicing of inteins was predicted based on first principles calculations. Using electronic structure methods, particularly density functional theory, the energies and structures were studied along the reaction pathway. Interesting observations include the catalyzing effect of the -NH_3^+ group of the N-terminal on the Cys+1 thiol group. In addition, the highly conserved N1-Cys1, N3-His10, C2-Asp5, C1-Asn8 and Cys+1 residues are involved in the splicing mechanism. An explanation for the lack of splicing activity with the C2-Asp5Gly mutation is discussed and the catalytic and essential

role of C2-Asp5 as an acid and a base is explained.

CHAPTER 6
CONCLUSIONS

6.1   Discussion

Intein-based technology is promising because a useful function is directly encoded into the intein sequence. An intein cuts itself away from the host protein and joins its left and right neighbors in a predictable way. Depending on sequence mutations and environmental stimuli, the function of inteins may be tuned, controlled, inhibited, or modified. Inteins may be used as a nano-sensor device, where the ligation of the exteins brings two fluorescent proteins (as exteins) together that are easily detected. Also, split inteins may be used for single molecule detection. In this case, a small molecule can trigger a detection device, which would combine N- and C-intein fragments which would then allow/catalyze splicing.

Because the enzyme-like intein is a component of the host protein (substrate) sequence, typical molecular inhibitors is not effective. For standard enzyme catalysis, the enzyme is usually a small molecule which can be replaced with a molecule that is chemically similar but not active and therefore inhibits the reaction. For inteins, the only evidence of chemical inhibition is from metal ions, which are unable to remain in the cell. From this complication, using inteins for biotechnological purposes has been difficult.

For this reason, it is important to understand the reaction mechanism on the atomic level. The role of protons is closely related to experimental crystal structures, random and site-directed mutagenesis, and NMR results. For example, NMR can compliment the crystal structure by understanding the protonation states of side chains. A complication with intein is that catalytic side chains such as Cys+1 are part of the exteins and are not included in the crystal structures. Mutagenesis experiments usually replace a normally active amino acid with something non-reactive, and the effect of this mutation on the reaction rate is measured. The splicing and cleavage mechanisms are important to understand on the atomic-level because this is how the intein will be used as a nanoswitch, molecular sensor, or reporting device.

6.2 Results

To investigate the reaction we have used first principles density functional theory methods to investigate the reaction mechanisms of inteins: N-terminal thioester formation, and transesterification, and C-terminal cleavage. Indeed, quantum mechanical methods allow for accurate prediction of energies related to biological systems, such as ionization potentials, electron affinities, binding energies, and energy barriers which could compliment experimental investigations. We have used these methods to analyze the structure and energetics of intein systems.

6.2.1 C-terminal cleavage

For the C-terminal cleavage mutant, we have proposed a mechanism for C-terminal cleavage based on C1-Asn7 cyclization and succinimide formation. Based on experimental results that showed the reaction to occur more rapidly in a low pH environment, our proposed mechanism takes this pH effect into account. We have also performed energy barrier calculations using QM, semi-empirical QM, implicit solvent QM, and QM/MM simulations to accurately recreate the reaction profile.

6.2.2 Splicing

In addition to C-terminal cleavage, we have studied the reaction steps of intein splicing, and found that the highly conserved N3-His10 plays an important role in accepting, then donating, a proton from the N1-Cys1. After N-terminal thioester formation, the polarization effect of the N-terminal $-NH_3^+$ group forces C-terminal Cys+1 to donate a proton to C2-Asp5. This step triggers transesterification, and the extra proton on C2-Asp5 catalyzes the C-terminal cleavage reaction in a way that mimics a low pH environment. With the C2-Asp5Gly mutation (CM), the N-terminal $-NH_3^+$ group is not stabilized, and returns to the neutral $-NH_2$ terminal group which is unable to catalyze the transesterification step, thus inhibiting the splicing reaction.

6.3 Future Research

Future research will include understanding specific mutations affecting intein splicing, and how these mutations may make the reaction occur more rapidly or stimuli dependent. For the *Mtu* recA intein, the mutation of C2-Asp5Gly decouples the N- and C-termini leading to isolated and rapid C-terminal cleavage. Also, the C-terminal Cys+1

residue may be mutated with the entire series of amino acids, and correlation with experiment tested.

Splicing may still occur with the mutations of the catalytic Cys to Ser/Thr residues, although with lower reaction rate due to the decreased acidity of the Ser side chain (-OH) compared to Cys (-SH) [4]. The effect of single and double Cys to Ser mutations can be studied with first principles calculations, and the intein splicing reaction may be tuned in order to control molecular docking related to inhibition or for a nanosensor device.

For the *Ssp* dnaE split intein, the effect of mutation in the extein (Tyr to Asp) mutation will be studied, especially on its effect on the protonation state of N3-His10, a supposed proton acceptor/donor for N-terminal thioester formation. One method to understand the effect of mutation is to calculate NMR chemical shifts for specific atoms based on a perturbed electrostatic environment due to mutation.

REFERENCES

[1] M. Belfort, B.L. Stoddard, D.W. Wood, and V. Derbyshire. *Homing Endonucleases and Inteins (Nucleic Acids and Molecular Biology)*. Springer, 2005.

[2] F.B. Perler, E.O. Davis, G.E. Dean, F.S. Gimble, W.E. Jack, N. Neff, C.J. Noren, J. Thorner, and M. Belfort. Protein splicing elements: inteins and exteins-a definition of terms and recommended nomenclature. *Nucleic Acids Research*, 22(7):1125–1127, 1994.

[3] K. Hiraga, V. Derbyshire, J.T. Dansereau, P. Van Roey, and M. Belfort. Minimization and stabilization of the Mycobacterium tuberculosis recA intein. *Journal of Molecular Biology*, 354(4):916–926, 2005.

[4] K. Shingledecker, S. Q. Jiang, and H. Paulus. Reactivity of the cysteine residues in the protein splicing active center of the Mycobacterium tuberculosis recA intein. *Archives of Biochemistry and Biophysics*, 375(1):138–144, 2000.

[5] H. Paulus. Protein splicing and related forms of protein autoprocessing. *Annual Review of Biochemistry*, 69(1):447–496, 2000.

[6] J. Miao, W. Wu, T. Spielmann, M. Belfort, V. Derbyshire, and G. Belfort. Single-step affinity purification of toxic and non-toxic proteins on a fluidics platform. *Lab on a Chip*, 5(3):248–253, 2005.

[7] M.R. Banki, L.A. Feng, and D.W. Wood. Simple bioseparations using self-cleaving elastin-like polypeptide tags. *Nature Methods*, 2(9):659–661, 2005.

[8] H. Paulus. Inteins as targets for potential antimycobacterial drugs. *Frontiers in Bioscience*, 8:s1157–s1165, 2003.

[9] H.D. Mootz and T.W. Muir. Protein splicing triggered by a small molecule. *Journal of the American Chemical Society*, 124(31):9044–9045, 2002.

[10] V. Muralidharan and T.W. Muir. Protein ligation: an enabling technology for the biophysical analysis of proteins. *Nature Methods*, 3:429–438, 2006.

[11] P. Van Roey, B. Pereira, Z. Li, K. Hiraga, M. Belfort, and V. Derbyshire. Crystallographic and mutational studies of Mycobacterium tuberculosis recA mini-inteins suggest a pivotal role for a highly conserved aspartate residue. *Journal of Molecular Biology*, 367(1):162–173, 2007.

[12] R. Mizutani, S. Nogami, M. Kawasaki, Y. Ohya, Y. Anraku, and Y. Satow. Protein-splicing reaction via a thiazolidine intermediate: crystal structure of the VMA1-derived endonuclease bearing the N and C-terminal propeptides. *Journal of Molecular Biology*, 316(4):919–929, 2002.

[13] B.W. Poland, M.Q. Xu, and F.A. Quiocho. Structural insights into the protein splicing mechanism of PI-SceI. *Journal of Biological Chemistry*, 275(22):16408–16413, 2000.

[14] T. Klabunde, S. Sharma, A. Telenti, W.R. Jacobs, and J.C. Sacchettini. Crystal structure of GyrA intein from Mycobacterium xenopi reveals structural basis of protein splicing. *Nature Structural Biology*, 5(1):31–36, 1998.

[15] X.Q. Duan, F.S. Gimble, and F.A. Quiocho. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell*, 89(4):555–564, 1997.

[16] K. Ichiyanagi, Y. Ishino, M. Ariyoshi, K. Komori, and K. Morikawa. Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. *Journal of Molecular Biology*, 300(4):889–901, 2000.

[17] P. Sun, S. Ye, S. Ferrandon, T.C. Evans, M.Q. Xu, and Z.H. Rao. Crystal structures of an intein from the split dnaE gene of Synechocystis sp PCC6803 reveal the catalytic model without the penultimate histidine and the mechanism of zinc ion inhibition of protein splicing. *Journal of Molecular Biology*, 353(5):1093–1105, 2005.

[18] M.A. Johnson, M.W. Southworth, T. Herrmann, L. Brace, F.B. Perler, and K. Wuthrich. NMR structure of a klba intein precursor from methanococcus jannaschii. *Protein Science*, 2007.

[19] D.W. Wood, W. Wu, G. Belfort, V. Derbyshire, and M. Belfort. A genetic system yields self-cleaving inteins for bioseparations. *Nature Biotechnology*, 17(9):889–892, 1999.

[20] M.W. Southworth, J. Benner, and F.B. Perler. An alternative protein splicing mechanism for inteins lacking an N-terminal nucleophile. *EMBO Journal*, 19(18):5019–5026, 2000.

[21] L.X. Chen, J. Benner, and F.B. Perler. Protein splicing in the absence of an intein penultimate histidine. *Journal of Biological Chemistry*, 275(27):20431–20435, 2000.

[22] K.V. Mills, J.S. Manning, A.M. Garcia, and L.A. Wuerdeman. Protein splicing of a Pyrococcus abyssi intein with a C-terminal glutamine. *Journal of Biological Chemistry*, 279(20):20685–20691, 2004.

[23] B. Pereira. Unpublished.

[24] D.W. Wood, V. Derbyshire, W. Wu, M. Chartrain, M. Belfort, and G. Belfort. Optimized single-step affinity purification with a self-cleaving intein applied to human acidic fibroblast growth factor. *Biotechnology Progess*, 16(6):1055–1063, 2000.

[25] M. Belfort, G. Belfort, V. Derbyshire, D. Wood, and W. Wu. Genetic system and self-cleaving inteins derived therefrom, bioseparations and protein purification employing same, and methods for determining critical, generalizable amino acid residues for varying intein activity. U.S. Patent #6,933,362, 2005.

[26] D.W. Wood. *Generation and application of a self-cleaving protein linker for use in single-step affinity fusion based protein purification.* PhD thesis, Rensselaer Polytechnic Institute, 2000.

[27] W. Wu, D.W. Wood, G. Belfort, V. Derbyshire, and M. Belfort. Intein-mediated purification of cytotoxic endonuclease I-TevI by insertional inactivation and pH-controllable splicing. *Nucleic Acids Research*, 30(22):4864–4871, 2002.

[28] P. Shemella, B. Pereira, Y. Zhang, P. Van Roey, G. Belfort, S. Garde, and S.K. Nayak. Mechanism for intein C-Terminal cleavage: A proposal from quantum mechanical calculations. *Biophysical Journal*, 92(3):847–853, 2007.

[29] P. Shemella, B. Pereira, P. Van Roey, G. Belfort, S. Garde, and S.K. Nayak. Controlling C-terminal cleavage rate of an intein through extein mutation: A quantum mechanical study. Submitted.

[30] B.M. Lew, K.V. Mills, and H. Paulus. Protein splicing in vitro with a semisynthetic two-component minimal intein. *Journal of Biological Chemistry*, 273(26):15887–15890, 1998.

[31] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review B*, 136(3B):864–871, 1964.

[32] W. Kohn and L.J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):1133–1138, 1965.

[33] R.P. Feynman. *Statistical mechanics.* Addison-Wesley, 1998.

[34] D.M. Ceperley and B.J. Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 45(7):566–569, 1980.

[35] R.M. Martin. *Electronic structure: Basic theory and practical methods.* Cambridge University Press, 2004.

[36] X. Wu, M.C. Vargas, S.K. Nayak, V. Lotrich, and G. Scoles. Towards extending the applicability of density functional theory to weakly bound systems. *Journal of Chemical Physics*, 115(19):8748–8757, 2001.

[37] M. van Schilfgaarde, T. Kotani, and S. Faleev. Quasiparticle self-consistent GW theory. *Physical Review Letters*, 96(22):226402, 2006.

[38] A.D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *Journal of Chemical Physics*, 98:5648–5652, 1993.

[39] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, 1988.

[40] J.P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244–13249, 1992.

[41] C.T. Lee, W.T. Yang, and R.G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988.

[42] B. Miehlich, A. Savin, H. Stoll, and H. Preuss. Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. *Chemical Physics Letters*, 157(3):200–206, 1989.

[43] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[44] X.D. Zhang, X.H. Zhang, and T.C. Bruice. A definitive mechanism for chorismate mutase. *Biochemistry*, 44(31):10443–10448, 2005.

[45] X.D. Zhang and T.C. Bruice. Reaction mechanism of guanidinoacetate methyltransferase, concerted or step-wise. *Proceedings of the National Academy of Sciences*, 103(44):16141–16146, 2006.

[46] C. Moeller and M.S. Plesset. Note on an approximation treatment for many-electron systems. *Physical Review*, 46(7):618–622, 1934.

[47] M. Head-Gordon, J.A. Pople, and M.J. Frisch. MP2 energy evaluation by direct methods. *Chemical Physics Letters*, 153(6):503–506, 1988.

[48] M.J. Frisch, M. Head-Gordon, and J.A. Pople. A direct MP2 gradient-method. *Chemical Physics Letters*, 166(3):275–280, 1990.

[49] M.J. Frisch, M. Head-Gordon, and J.A. Pople. Semidirect algorithms for the MP2 energy and gradient. *Chemical Physics Letters*, 166(3):281–289, 1990.

[50] W.J. Hehre, R. Ditchfield, and J.A. Pople. Self-consistent molecular orbital methods. XII. Further extensions of Gaussiantype basis sets for use in molecular orbital studies of organic molecules. *Journal of Chemical Physics*, 56(5):2257–2261, 1972.

[51] A.D. McLean and G.S. Chandler. Contracted Gaussian basis sets for molecular calculations. 1. Second row atoms, Z=11-18. *Journal of Chemical Physics*, 72(10):5639–5648, 2006.

[52] J.J.P. Stewart. Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry*, 10(2):209–220, 1989.

[53] D.M. Ferguson, I.R. Gould, W.A. Glauser, S. Schroeder, and P.A. Kollman. Comparison of ab initio, semiempirical, and molecular mechanics calculations for the conformational–analysis of ring–systems. *Journal of Computational Chemistry*, 13(4):525–532, 1992.

[54] G.I. Csonka. Analysis of the core-repulsion functions used in AM1 and PM3 semiempirical calculations: conformational analysis of ring systems. *Journal of Computational Chemistry*, 14(8):895–898, 1993.

[55] S. Miertus, E. Scrocco, and J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, 55(11):117–129, 1981.

[56] M.T. Cances, B. Mennucci, and J. Tomasi. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *Journal of Chemical Physics*, 107(8):3032–3041, 1997.

[57] F. Maseras and K. Morokuma. IMOMM: A new integrated ab initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *Journal of Computational Chemistry*, 16(9):1170–1179, 1995.

[58] T. Vreven, K. Morokuma, O. Farkas, H.B. Schlegel, and M.J. Frisch. Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints. *Journal of Computational Chemistry*, 24(6):760–769, 2003.

[59] M.A. Thompson and G.K. Schenter. Excited states of the Bacteriochlorophyll b dimer of Rhodopseudomonas viridis: A QM/MM study of the photosynthetic reaction center that includes MM polarization. *Journal of Physical Chemistry*, 99(17):6374–6386, 1995.

[60] J.C. Schoneboom, H. Lin, N. Reuter, W. Thiel, S. Cohen, F. Ogliaro, and S. Shaik. The elusive oxidant species of cytochrome P450 enzymes: Characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. *Journal of the American Chemical Society*, 124(27):8142–8151, 2002.

[61] J.L. Gao and D.G. Truhlar. Quantum mechanical methods for enzyme kinetics. *Annual Review of Physical Chemistry*, 53:467–505, 2002.

[62] Q. Cui, M. Elstner, and M. Karplus. A theoretical analysis of the proton and hydride transfer in liver alcohol dehydrogenase (LADH). *Journal of Physical Chemistry B*, 106(10):2721–2740, 2002.

[63] M. Torrent, T. Vreven, D.G. Musaev, K. Morokuma, O. Farkas, and H.B. Schlegel. Effects of the protein environment on the structure and energetics of active sites of metalloenzymes. ONIOM study of methane monooxygenase and ribonucleotide reductase. *Journal of the American Chemical Society*, 124(2):192–193, 2002.

[64] C. Dellago, P.G. Bolhuis, F.S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *Journal of Chemical Physics*, 108(5):1964–1977, 1998.

[65] K.K. Irikura and D.J. Frurip. *Computational thermochemistry: Prediction and estimation of molecular thermodynamics*. American Chemical Society, 1998.

[66] J.W. Ochterski. *Thermochemistry in Gaussian*. Gaussian Inc., 2000.

[67] T. Geiger and S. Clarke. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *Journal of Biological Chemistry*, 262(2):785–794, 1987.

[68] S. Capasso, L. Mazzarella, G. Sorrentino, G. Balboni, and A.J. Kirby. Kinetics and mechanism of the cleavage of the peptide bond next to asparagine. *Peptides*, 17(6):1075–1077, 1996.

[69] F. Catanzano, G. Graziano, S. Capasso, and G. Barone. Thermodynamic analysis of the effect of selective monodeamidation at asparagine 67 in ribonuclease A. *Protein Science*, 6(8):1682–1693, 1997.

[70] B. Peters and B.L. Trout. Asparagine deamidation: pH-dependent mechanism from density functional theory. *Biochemistry*, 45(16):5384–5392, 2006.

[71] Y. Ding, M.Q. Xu, I. Ghosh, X.H. Chen, S. Ferrandon, G. Lesage, and Z.H. Rao. Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. *Journal of Biological Chemistry*, 278(40):39133–39142, 2003.

[72] E.J. Milner-White. The partial charge of the nitrogen atom in peptide bonds. *Protein Science*, 6(11):2477–2482, 1997.

[73] D. Voet and J.G. Voet. *Biochemistry*. Wiley, 1995.

[74] J. Trylska, P. Grochowski, and J.A. McCammon. The role of hydrogen bonding in the enzymatic reaction catalyzed by HIV-1 protease. *Protein Science*, 13(2):513–528, 2004.

[75] J.P. Krug, P.L.A. Popelier, and R.F.W. Bader. Theoretical study of neutral and of acid and base-promoted hydrolysis of formamide. *Journal of Physical Chemistry*, 96(19):7604–7616, 1992.

[76] R.K. Eppler, R.S. Komor, J. Huynh, J.S. Dordick, J.A. Reimer, and D.S. Clark. Water dynamics and salt-activation of enzymes in organic media: Mechanistic implications revealed by NMR spectroscopy. *Proceedings of the National Academy of Sciences*, 103(15):5706–5710, 2006.

[77] E.O. Davis, S.G. Sedgwick, and M.J. Colston. Novel structure of the recA locus of Mycobacterium tuberculosis implies processing of the gene product. *Journal of Bacteriology*, 173(18):5653–5662, 1991.

[78] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[79] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, and H.J.C. Berendsen. GROMACS: Fast, flexible, and free. *Journal of Compututational Chemistry*, 26(16):1701–1718, 2005.

[80] P. Shemella and S.K. Nayak. Unpublished.

[81] A.E. Reed, L.A. Curtiss, and F. Weinhold. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chemical Reviews*, 88(6):899–926, 1988.

[82] Y.W. Bai, J.S. Milne, L. Mayne, and S.W. Englander. Primary structure effects hydrogen exchange. *PROTEINS: Structure, Function, and Genetics*, 17(1):75–86, 1993.

[83] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[84] B. Pereira, S. Jain, and S. Garde. Quantifying the protein core flexibility through analysis of cavity formation. *Journal of Chemical Physics*, 124:074704, 2006.

[85] R.S. Mulliken. Electronic population analysis on LCAO [single bond] MO molecular wave functions. I. *Journal of Chemical Physics*, 23(10):1833–1840, 1955.

[86] B.M. Lew and H. Paulus. An in vivo screening system against protein splicing useful for the isolation of non-splicing mutants or inhibitors of the RecA intein of Mycobacterium tuberculosis. *Gene*, 282(1-2):169–177, 2002.

[87] K.V. Mills and H. Paulus. Reversible inhibition of protein splicing by zinc ion. *Journal of Biological Chemistry*, 276(14):10832–10838, 2001.

[88] B. Dassa, G. Amitai, J. Caspi, O. Schueler-Furman, and S. Pietrokovski. Trans protein splicing of cyanobacterial split inteins in endogenous and exogenous combinations. *Biochemistry*, 46(1):322–330, 2007.

[89] N.D. Clarke. A proposed mechanism for the self-splicing of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 91(23):11084–11088, 1994.

[90] G. Amitai and S. Pietrokovski. Fine-tuning an engineered intein. *Nature Biotechnology*, 17(9):854–855, 1999.

[91] H. Iwai, S. Zuger, J. Jin, and P.H. Tam. Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. *FEBS Letters*, 580(7):1853–1858, 2006.

[92] A.E. Gorbalenya. Non-canonical inteins. *Nucleic Acids Research*, 26(7):1742, 1998.

[93] B. Dassa, H. Haviv, G. Amitai, and S. Pietrokovski. Protein splicing and auto-cleavage of bacterial intein-like domains lacking a C'-flanking nucleophilic residue. *Journal of Biological Chemistry*, 279(31):32001–32007, 2004.

[94] G. Amitai, B. Dassa, and S. Pietrokovski. Protein splicing of inteins with atypical glutamine and aspartate C-terminal residues. *Journal of Biological Chemistry*, 279(5):3121–3131, 2004.

[95] R.C. Stephenson and S. Clarke. Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins. *Journal of Biological Chemistry*, 264(11):6164–6170, 1989.
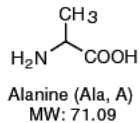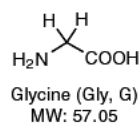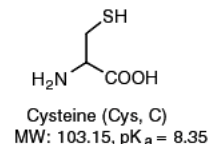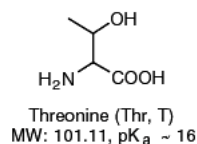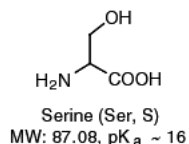
APPENDIX A
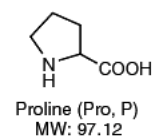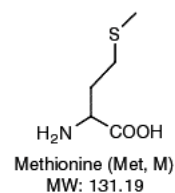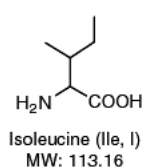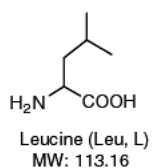# Amino Acid Chart

Twenty natural amino acids (source http://www.neb.com).

Small

Glycine (Gly, G)
MW: 57.05

Alanine (Ala, A)
MW: 71.09

Nucleophilic

Serine (Ser, S)
MW: 87.08, $pK_a \sim 16$

Threonine (Thr, T)
MW: 101.11, $pK_a \sim 16$

Cysteine (Cys, C)
MW: 103.15, $pK_a = 8.35$

Hydrophobic

Valine (Val, V)
MW: 99.14

Leucine (Leu, L)
MW: 113.16

Isoleucine (Ile, I)
MW: 113.16

Methionine (Met, M)
MW: 131.19

Proline (Pro, P)
MW: 97.12

Aromatic

Phenylalanine (Phe, F)
MW: 147.18

Tyrosine (Tyr, Y)
MW: 163.18

Tryptophan (Trp, W)
MW: 186.21

Acidic

Aspartic Acid (Asp, D)
MW: 115.09, $pK_a = 3.9$

Glutamic Acid (Glu, E)
MW: 129.12, $pK_a = 4.07$

Amide

Asparagine (Asn, N)
MW: 114.11

Glutamine (Gln, Q)
MW: 128.14

Basic

Histidine (His, H)
MW: 137.14, $pK_a = 6.04$

Lysine (Lys, K)
MW: 128.17, $pK_a = 10.79$

Arginine (Arg, R)
MW: 156.19, $pK_a = 12.48$

Intein sequences, motifs, and numbering scheme

Old block	New block	Position on block	*Mtu* recA	WT+EN	*Mtu* recA 110△383	*Mtu* recA 94△403	*Ssp*6803 dnaE	*Ssp*6803 dnaE
A	N1	1	Cys	1	1	1	Cys	1
A	N1	2	Leu	2	2	2	Leu	2
A	N1	3	Ala	3	3	3	Ser	3
A	N1	4	Glu	4	4	4	Phe	4
A	N1	5	Gly	5	5	5	Gly	5
A	N1	6	Thr	6	6	6	Thr	6
A	N1	7	Arg	7	7	7	Glu	7
A	N1	8	Ile	8	8	8	Ile	8
A	N1	9	Phe	9	9	9	Leu	9
A	N1	10	Asp	10	10	10	Thr	10
A	N1	11	Pro	11	11	11	Val	11
A	N1	12	Val	12	12	12	Glu	12
A	N1	13	Thr	13	13	13	Tyr	13
-	N2	1	His	17	17	17	Leu	16
-	N2	2	Arg	18	18	18	Pro	17
-	N2	3	Ile	19	19	19	Ile	18
-	N2	4	Glu	20	20	20	Gly	19
-	N2	5	Asp	21	21	21	Lys	20
-	N2	6	Val	22	22	22	Ile	21
-	N2	7	Val	23	23	23	Val	22
-	N2	8	Asp	24	24	24	Ser	23
B	N3	1	Gly	64	64	64	Gly	63
B	N3	2	Ala	65	65	65	Ser	64
B	N3	3	Ile	66	66	66	Val	65
B	N3	4	Val	67	67	67	Ile	66
B	N3	5	Trp	68	68	68	Arg	67
B	N3	6	Ala	69	69	69	Ala	68
B	N3	7	Thr	70	70	70	Thr	69
B	N3	8	Pro	71	71	71	Ser	70
B	N3	9	Asp	72	72	72	Asp	71
B	N3	10	His	73	73	73	His	72
B	N3	11	Lys	74	74	74	Arg	73
B	N3	12	Val	75	75	75	Phe	74
B	N3	13	Leu	76	76	76	Leu	75
B	N3	14	Thr	77	77	77	Thr	76
-	N4	1	Trp	81	81	81	Leu	81
-	N4	2	Arg	82	82	82	Leu	82
-	N4	3	Ala	83	83	83	Ala	83
-	N4	4	Ala	84	84	84	Ile	84
-	N4	5	Gly	85	85	85	Glu	85
-	N4	6	Glu	86	86	86	Glu	86
-	N4	7	Leu	87	87	87	Ile	87
-	N4	8	Arg	88	88	88	Phe	88
-	N4	9	Lys	89	89	89	Ala	89
-	N4	10	Gly	90	90	90	Arg	90
-	N4	11	Asp	91	91	91	Gln	91
-	N4	12	Arg	92	92	92	Leu	92
-	N4	13	Val	93	93	93	Asp	93
-	N4	14	Ala	94	94	94	Leu	94
-	N4	15	Gln	95	95	-	Leu	95
-	N4	16	Pro	96	96	-	Thr	96
F	C2	1	Ala	418	146	119	Gln	136
F	C2	2	Arg	419	147	120	Arg	137
F	C2	3	Thr	420	148	121	Ile	138
F	C2	4	Phe	421	149	122	Phe	139
F	C2	5	Asp	422	150	123	Asp	140
F	C2	6	Leu	423	151	124	Ile	141
F	C2	7	Glu	424	152	125	Gly	142
F	C2	8	Val	425	153	126	Leu	143
F	C2	9	Glu	426	154	127	Pro	144
F	C2	10	Glu	427	155	128	Gln	145
F	C2	11	Leu	428	156	129	Asp	146
F	C2	12	His	429	157	130	His	147
F	C2	13	Thr	430	158	131	Asn	148
F	C2	14	Leu	431	159	132	Phe	149
F	C2	15	Val	432	160	133	Leu	150
G	C1	1	Ala	433	161	134	Ala	152
G	C1	2	Glu	434	162	135	Asn	153
G	C1	3	Gly	435	163	136	Gly	154
G	C1	4	Val	436	164	137	Ala	155
G	C1	5	Val	437	165	138	Ile	156
G	C1	6	Val	438	166	139	Ala	157
G	C1	7	His	439	167	140	Ala	158
G	C1	8	Asn	440	168	141	Asn	159
G	C1	1	Cys	441	169	142	Cys	160